sciendo

studia humana
QUARTERLY JOURNAL

Studia Humana
Volume 8:1 (2019)

Contents

sciendo

studia humana
QUARTERLY JOURNAL

Studia Humana
Volume 8:1 (2019), pp. 3—18
DOI: 10.2478/sh-2019-0001

AI Case Studies: Potential for Human Health, Space Exploration and Colonisation and a Proposed Superimposition of the Kubler-Ross Change Curve on the Hype Cycle

Matthew Williams

Universe Today, Canada

e-mail: houseofwilliams@yahoo.ca

Martin Braddock

Sherwood Observatory,
Sutton-in-Ashfield, England, United Kingdom

e-mail: projects@sherwood-observatory.org.uk

Abstract:
The development and deployment of artificial intelligence (AI) is and will profoundly reshape human society, the culture and the composition of civilisations which make up human kind. All technological triggers tend to drive a hype curve which over time is realised by an output which is often unexpected, taking both pessimistic and optimistic perspectives and actions of drivers, contributors and enablers on a journey where the ultimate destination may be unclear. In this paper we hypothesise that this journey is not dissimilar to the personal journey described by the Kubler-Ross change curve and illustrate this by commentary on the potential of AI for drug discovery, development and healthcare and as an enabler for deep space exploration and colonisation. Recent advances in the call for regulation to ensure development of safety measures associated with machine-based learning are presented which, together with regulation of the rapidly emerging digital after-life industry, should provide a platform for realising the full potential benefit of AI for the human species.
Keywords: artificial intelligence, regulation, healthcare, space exploration, digital afterlife.

1. Introduction

AI is the intelligence demonstrated by machines in contrast to natural intelligence which is displayed by human beings and other animals. Weak AI, also known as narrow AI, is artificial

intelligence that is focused on one narrow task, for example satellite navigation. In contrast to weak AI, strong AI, also defined as broad AI, artificial general intelligence (AGI) or sentient AI is a machine with consciousness, sentience and mind and with the ability to apply intelligence to any problem, rather than just one specific task [1]. Today, all systems that use AI are operating as a weak AI focused on a narrowly defined specific problem, however, such is the advance in the field, the possibility of developing stronger if not strong AI may become a reality as computing power continues to increase and digital data is used in reinforced learning algorithms.

Alan Turing [2] defined intelligence in a very simple way as a question of conversation. He proposed that if a machine can answer any question on any subject asked of it using the same words and tone that a human being can understand, then the machine is called intelligent. This definition is known as the Turing test.

The objectives of this paper were to stimulate thinking for the assessment of the potential of AI in two areas of science and to support a philosophical argument that progression along the Kubler-Ross change curve can match the profile of new technologies defined by the hype cycle. We have chosen the areas of drug discovery and development and their application to the rapidly changing field of human healthcare and the prospects for space exploration and colonisation, both today as current examples of weak AI. However, the potential for strong AI, despite a distant goal should not be underestimated, indeed it may form part of the future for deep space exploration and colonisation of exoplanets outside of the Solar system.

2. Definition of Life

There are many definitions of life and we capture three which represent the most consistent:

"The condition that distinguishes animals and plants from inorganic matter, including the capacity for growth, reproduction, functional activity, and continual change preceding death" [3].

"The property or quality that distinguishes living organisms from dead organisms and inanimate matter, manifested in functions such as metabolism, growth, reproduction, and response to stimuli or adaptation to the environment originating from within the organism" [4].

"Life, living matter and, as such, matter that shows certain attributes that include responsiveness, growth, metabolism, energy transformation, and reproduction" [5].

The definition of life, as for The Turing test is anthropomorphic in nature. As the definition is usually derived from biologists, it will by inference relate traits identifiable within the life forms studied, which in turn have a common origin. This situation establishes a bias on both the observation and therefore the output. We consider parameters which support the classical definition of life today and extend the terms into a consideration of a possible future definition, making some key observations and asking several questions (Table 1). We note that that there are similarities between parameters which support the human definition of life and those which could be applied to AI.

As we begin to become less constrained by the current boundaries of human and machine intelligence and perhaps venture towards elective human enhancement in future, rather than therapeutic enhancement today, our traditional definition of life will require revisiting. Perhaps a more fitting definition of life could be: "A system which is self-sustained and able to learn and adapt to environmental input".

Table 1. Parameter Definition of Life Today and a Proposed Future Definition

Parameter	Today	Future	Key observations and questions
Metabolism	Classical definition of life today which differentiates animate from inanimate systems, requiring homeostatic processes to maintain longevity	AI consumes power to drive function and may autoregulate power consumption in future.	Could AI develop a more efficient power system autonomously? Nutrition and excretion components of metabolism are analogous to memory resources and file space used (nutrition) and discarded (excretion).
Growth		An algorithm can be programmed to grow. AI does not undergo cell division but could be programmed to replicate at a pre-defined point.	Viruses and AI do not have a cell cycle as viruses need to infect host cells to replicate. Does the definition of AI resemble that of a virus?
Life cycle		AI may have a life cycle through platforms such as Alexa, Siri and the Replika app.	The life cycle may progress or replicate via access of next generation platforms to digital data. Will this continue to be with the permission of the individual or organisation who generate the data?
Reproduce		An algorithm can be programmed to reproduce.	A self-replicating neural network has been created and a population of AI agents may be able self-improve through a form of natural selection [6].
Respond to stimuli		Emotional intelligence is rapidly evolving in cameras to be able to detect emotional and facial expression metrics. The race is on to recognise individual body language and tone of voice and provide personalised interaction.	It is likely that AI will be able to sense some forms of human emotive behaviour [7] which may be tuned to accommodate difference in cultures. How will this ability be used?
Adapt to environment		Robotic AI is being developed which may adapt to the environment to overcome obstacles.	It is likely that AI will enable drones or terrestrial robots to adapt to hostile environment such as earthquake zones or other areas hazardous to human intervention [8]. Could and should this be

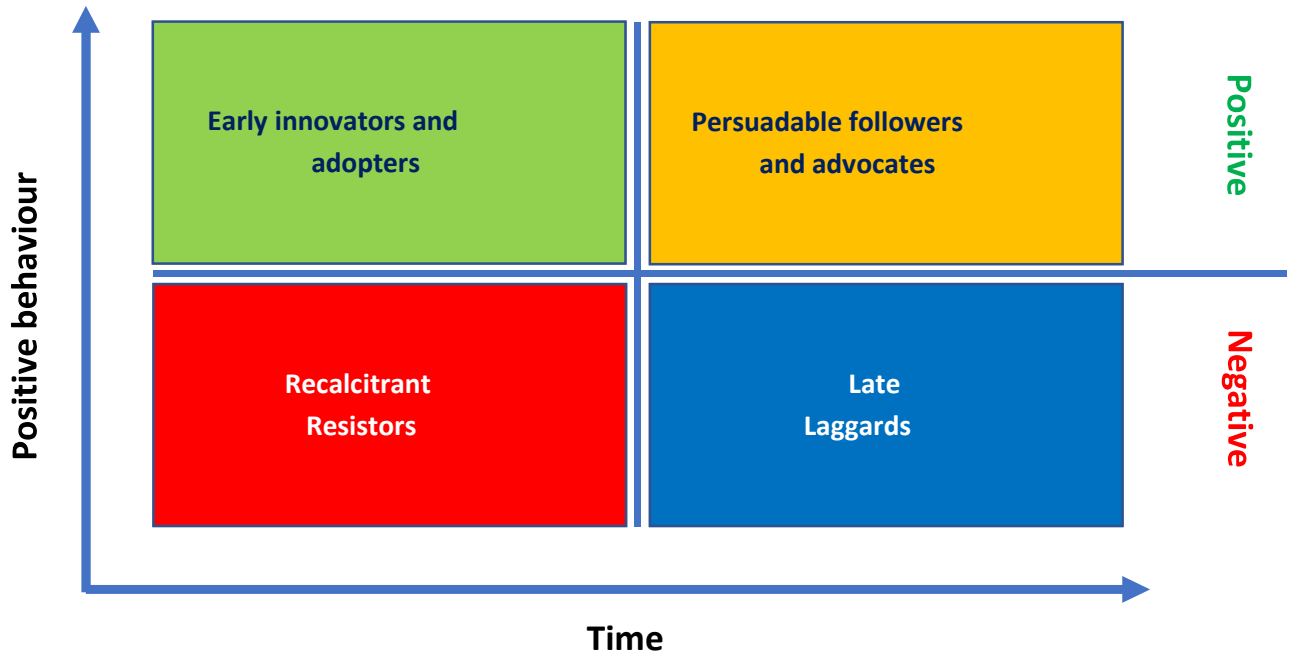
			extended into areas of conflict?
Evolve		AI has been shown to evolve to solve problems.	Neuroevolution, where neural networks are optimized through evolutionary algorithms, is an effective method to train deep neural networks for reinforcement learning (RL) problems [9-13]. The use of simple genetic algorithms has surprisingly outperformed state-of-the-art RL algorithms.

3. Hype Cycle

The hype cycle is a graphical representation of the maturity, adoption and application in the real world of specific technologies. It is a branded product of Gartner, a US based company dealing in consultancy and advisory activities applying information technology to the assessment of emerging trends in many technologies [14]. It has been particularly useful in the mapping and assessment of the myriad of activities which fall under the category of AI [15] and may be summarised in Amara's law which states that: "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run" [16].

The generation of innovation and adoption of new technology will naturally be met by a variety of human response characteristics and four such categories may be represented which will drive, influence and enable both the development and deployment of new technology (Figure 1).

Figure 1



The figure may be considered in two parts; an upper positive level and a lower negative level. On the upper level. ‘early innovators and adopters’ are those individuals who are in the main optimistic, are entrepreneurial and have either the experience or persuasive power to be highly influential to others, especially technical experts whose skills are required. The second quadrant of positive individuals are the ‘persuadable followers and advocates’ who are essential catalysts for change and are likely a larger group of people. The lower level may be comprised of two quadrants. The first are the ‘late laggards’, who either through apathy or eventual realisation join their colleagues in progressing a concept. The last group are the ‘recalcitrant resistors’, who tend to be in the minority and who are resistant to new opportunities and can even be damaging to those ‘persuadable followers and advocates’. This is particularly relevant when the nature of human optimism is addressed but first we should consider the human response to change.

4. Kubler-Ross Change Curve and Criticism

The change curve was originally proposed in 1969 by Elisabeth Kubler-Ross to illustrate and help people cope with and manage a personal terminal illness or an illness associated with a close relative or friend [17]. It is now widely used to manage dramatic change and perceived crisis, especially in larger organisations where staff experience sudden news, often associated with major restructuring [18]. By understanding both personal and personnel’s response to change, understanding a person’s position on the change curve may help manage the situation by retaining perspective and objectivity and ultimately bringing forward a transition of leaders and their followers to the new normale. It is not a tool without criticism and indeed in the example of coping with personal grief due to death, evidence substantiating complete transition through the change curve may be lacking [19] leading to caution in the belief that everybody completely adapts to change over time [20]. As with all change, an inherent pessimism or optimism plays a great role in retaining both personal motivation and objectivity and also in signalling the positive messages associated with change acceptance to others with whom individuals have either direct or indirect contact and communication.

5. Change-hype and Despair Versus Hype and Hope

Although our consideration of the hype cycle and change curve is only briefly addressed, many examples across multiple industries support both as models of human response and reaction and over time, both models allow eventual establishment of a *status quo* even though this may be transitory. We will consider two further factors, one example for each of the hype cycle and the change curve. The first, specific to AI is known as the AI winter and the avoidance of a second occurrence [21], [22]. The AI winter corresponds to the period at the trough of disillusionment [14] where, in this case, investor confidence and expectation, public interest and promise reached rock bottom and funding and commitment to AI technologies was drastically reduced. Many of the reasons which brought on the onset of past AI winters involved over-expectation and promise where there was and remains some concern that expectations will not be met [23] or that AI will become sentient [24]. To help put the opportunity presented by AI in perspective, a recent European Parliamentary Report exposes some of the myths and calls for a global charter to help maximise the benefits of a technology often misunderstood and misrepresented to the general public [25]. This misunderstanding is further illustrated by a report describing public perception and pessimism in long-term trends in the development of AI, though encouragingly finds optimism in a future of AI in healthcare and education [26]. This forms a segue way into the next factor which relates to personal optimism. It has been known for some years that there is a neural basis for optimism, where both optimism and pessimism are associated with different parts of the brain’s cerebral hemisphere, pessimism with neurological processes in the right hemisphere and conversely, optimism with neurological processes in the left hemisphere [27], [28]. Moreover, there is evidence

that optimists already preconceive a situation where there is probability of a future successful outcome [29] and these *a priori* beliefs will likely be present in individuals who reside in the upper left quadrant of Figure 1. Furthermore, unrealistic optimism though perhaps useful in conditions of high adversity, is a potential danger, particularly in the context of decision making on investment of personal or others' funds and time. Clearly individual psyche may have a major influence on esteem and personal health and well-being and directly impact the behaviour and esteem of others, particularly from a position of influence. We notice that there is a similarity in the profile of the hype cycle and the Kubler-Ross change curve and have superimposed both curves. Accepting that the axes are arbitrary and that the magnitude is for illustration purposes only, we have considered what the Kubler-Ross change curve may be for a person who is neither a natural pessimist or optimist (Figure 2a) and a natural optimist (Figure 2b).

Figure 2a

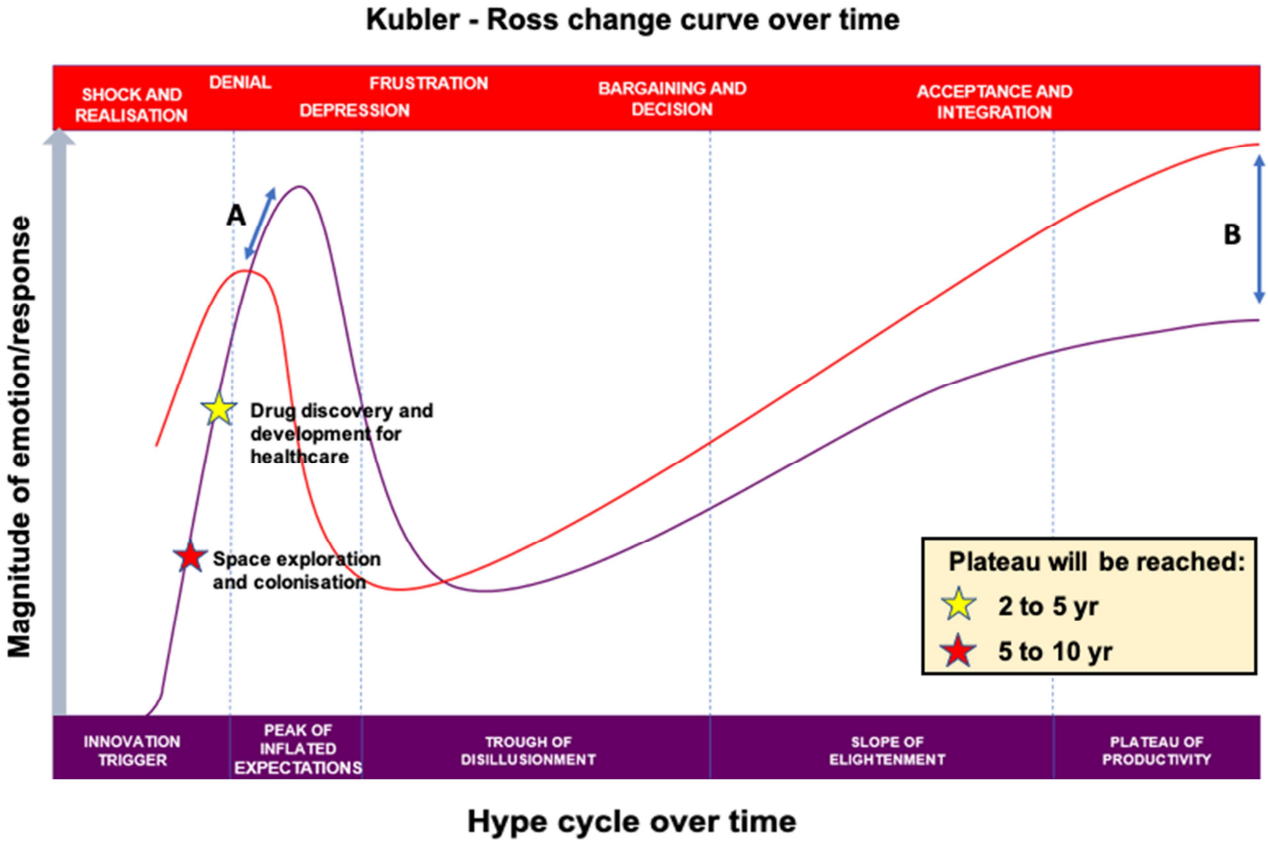
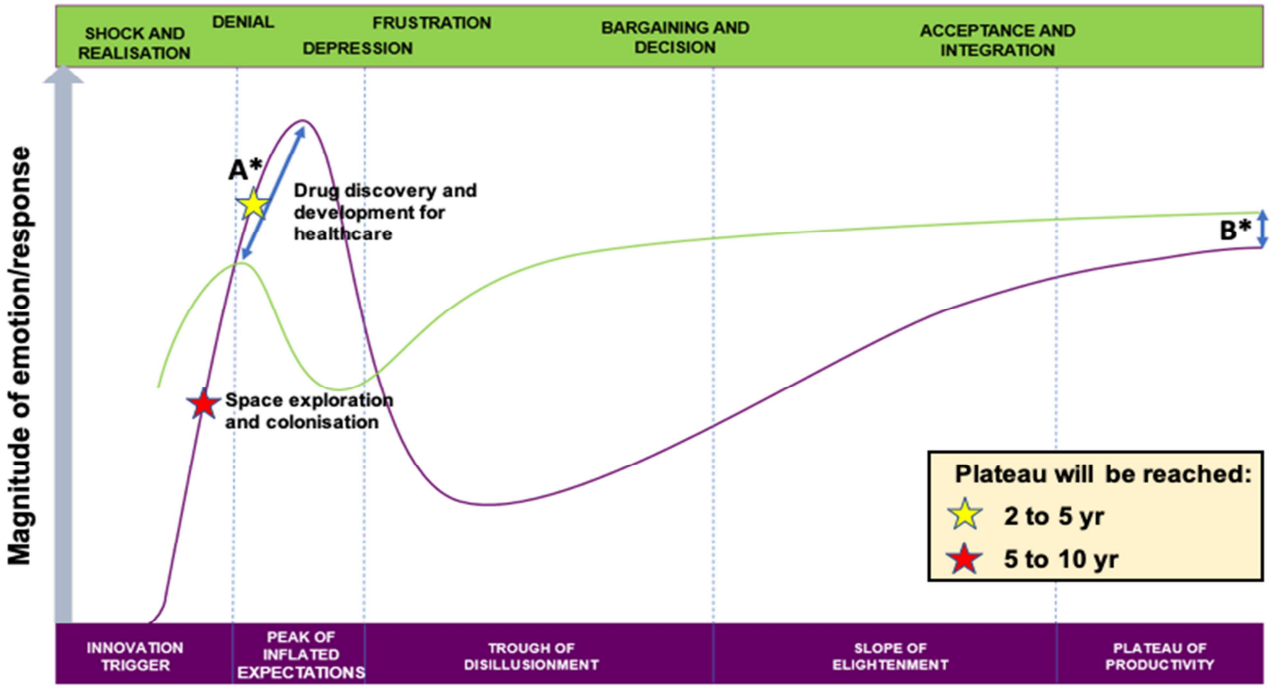


Figure 2b



Hype cycle over time

A natural pessimist, or a person who is neither a pessimist nor an optimist may experience a rapid and steep shock behaviour where the peak of their response precedes the peak of expectation with a difference (A) and over time their acceptance and integration level is higher than at the start of their entry on the change curve with a difference magnitude (B). Compare this profile with that of a person predisposed to optimism, where their emotional experience on the change curve is blunted and flatter such that A* > A and the difference magnitude at the end of the cycle and curve is such that B* < B. It should be noted that this is an illustrative proposal and that further qualitative work is needed to verify this concept which attempts to accommodate the natural predisposition to pessimism or optimism with the realisation and eventual acceptance of new technology.

6. Case Study One – AI in Drug Discovery, Development and Healthcare Today

In 2016, a study by the Tufts Centre for the Study of Drug Development estimated the overall cost of delivering a new approved and marketed medicine at $2.6Bn [30]. Reasons underlying productivity decline are multi-factorial and comprise contributions from basic non-clinical and translational science, clinical efficacy and safety, regulatory and commercial issues, together with the need to tackle increasingly challenging areas of human disease where the pathophysiology is often heterogeneous. Significant advances in screening [31], use of antibodies [32], [33] and generation of new modalities for targets previously thought as intractable [34], [35], phenome technologies applied to large samples sets and small volume sample size [36] mean that it is now possible to generate very large data sets designed to assist the selection of candidate drugs and their progression through lengthy and costly clinical trials. One critical part of the drug discovery process which underpins all downstream further drug development in both non-clinical and clinical phases is the physical laboratory-based manual handling of potential new medical entities and to match appropriately designed drugs to the genotype and phenotype of the patient. Drug discovery is often described using the metaphor of finding a needle in a haystack. In this case, the haystack comprises the order of 10^{60} - 10^{100} synthetically feasible molecules [37], out of which a compound needs to be identified which either satisfies all the standard criteria for a molecule with drug like properties or can form a lead for further optimisation. Either way, the fraction of the total number of molecules

which can be physically synthesised, let alone tested *in vitro* is very small and algorithm based virtual *de novo* design of molecules may produce a more restricted and manageable chemical library for laboratory-based work to commence or further progress.

The rapidly expanding area of machine-based learning artificial intelligence has been used to 'teach' computers the basic principles of drug design from the foundation. AI approaches to rational drug design and to predicting drug toxicology have been proposed for many years [38] and numerous approaches have been taken and reported which describe varying degrees of success [39-42], with some questions remaining over the ability of AI to reproduce true chemical diversity, a prerequisite for exploitation of chemical properties aligned to delivering novelty [43]. More recently, an algorithm known as Reinforcement Learning for Structural Evolution, or ReLeaSE, is an algorithm and computer program that comprises two neural networks [44]. The networks may be regarded as a trainer and a learner network, where the trainer employs the syntax and linguistic rules for the language of chemical structures for approximately 1.7 million known biologically active molecules. This machine-based learning approach has been successful in designing compounds that fall into two broad and opposing classes of differing melting temperatures, biased towards a range of lipophilicity and with differing values of pIC_{50} directed against the Janus 2 non-receptor tyrosine kinase. The system can be extended to multi-parameter optimization of compound properties in a concurrent manner and it may be possible to compress the classical lead identification and optimization phases, build into desired targets pharmacophores which may afford, for example radioprotection as standard while optimising potency, selectivity, solubility, and DMPK parameters associated with drug-likeness.

For both a pessimist and an optimist we place drug discovery, development and healthcare on the upward innovation trigger curve expecting the plateau to be reached within the next 2 to 5 years with the optimist higher up the curve.

7. Tomorrow

Given the driver to improve overall costs associated with research and development for new medicines and the potential of AI, there are very high expectations on the technology to shorten discovery cycle times and to align new drug design with digital data that is, or will be collected on individuals to ensure the 'right drug to the right patient taken in the right way' [45]. A rapid increase in the number of collaborations of AI technology companies with key players in the biopharmaceutical industry has occurred particularly over the last 2-3 years [46]. In addition to drug hunting and alignment to patient needs, AI also has a potential role in maximising the value of established drugs or discontinued drugs to be repurposed for the treatment of human disease for which they were not originally intended. Known as drug repurposing, this has been predominantly a serendipitous approach which has progressed well along the hype cycle having generally delivered less than expected and having been faced with numerous challenges [47]. An algorithm assisted way of predicting on-target activity in multiple disease indications and exploiting off-target activity in the same, whilst minimising both on and off-target toxicology may allow drug developers to more rapidly assemble a candidate list for clinical testing further informed by the collection of patient data. However, the potential for the contribution of AI to over promise has been recognised as 'the storm before the calm' [48].

In February 2019, it was announced that researchers have developed the largest virtual library of molecules which will soon contain over a billion molecules in a free publicly accessible pharmacology platform called ZINC and have shown that it can identify new chemotypes, some of which have very good potency and selectivity directed at their molecular target [49]. This is potentially a paradigm shift in AI assisted drug discovery and it remains to be seen whether this technology can be exploited to provide faster routes to drug candidate selection for clinical testing than the current industry standard.

One further area which has emerged into prominence is the field of AI assisted clinical diagnosis and management of patients, termed by Eric Topol as Deep Medicine, which may be a true partnering of AI with human intelligence and experience [50]. There can be no doubt that AI is providing substantial benefit to healthcare as the number of AI based approvals from the Federal Drug Administration is increasing and the number of indications reporting favourable predictive power for AI is also increasing. However, the paper reports that: "The state of AI hype has far exceeded the state of AI science" and provides a balanced assessment of the limitations and challenges and clear recommendations for future consideration which envisage AI as a vital *assistant* to the physician and not a *replacement* for him or her.

8. Case Study Two – AI in Space Exploration and Colonization Today

The development of machine learning, and artificial neural networks is also having a significant impact on space exploration. Already, space agencies like the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA) are looking to AI for assistance with data collection, data analysis, mission planning, guidance and research target selection. The first-ever case of AI being used to assist with space exploration is the Deep Space 1 probe, a technology demonstrator tasked with conducting a flyby of an asteroid (9969 Braille) and a comet (Borrelly) in 1998. This mission used an AI algorithm called Remote Agent [51], which can plan activities and diagnosing failures on-board. Since then, AI has played a major role in assisting with Earth observations and astronomical research. Examples include the Earth Observer 1 (EO-1) satellite [52], which relies on AI systems to optimize its analysis and response to natural disasters; and the Sky Image Cataloguing and Analysis Tool (SKICAT) [53], which relies on AI to assist with the classification of objects discovered by the second Palomar Sky Survey.
In addition, a team of astronomers used machine learning to sort through data gathered by the Kilo Degree Survey (KiDS) to identify 56 new possible gravitational lenses [54]. As next-generation telescopes commence operations, researchers plan to use AI to find patterns and correlations by systematic interrogation of vast amounts of gathered data. In terms of exploration, AI is already playing an important role. This includes the Autonomous Exploration for Gathering Increased Science (AEGIS), which provides automated targeting or remote sensing instruments on the Spirit and Opportunity rovers [55]; and the Planetary Instrument for X-ray Lithochemistry (PIXL), an autonomous instrument developed for the *Mars 2020* rover that is designed to examine fine scale chemical variations in rocks and soils on planetary surfaces [56] - a key indicator of past (or present) life. On PIXL, John Leif Jørgensen from DTU Space in Denmark said:

> PIXL's microscope is situated on the rover's arm and needs to be placed 14 millimetres from what we want it to study. That happens thanks to several cameras placed on the rover. It may sound simple, but the handover process and finding out exactly where to place the arm can be likened to identifying a building from the street from a picture taken from the roof. This is something that AI is eminently suited for [57].

Looking ahead, further developments in the field of AI are expected to have an even more significant impact. With applications ranging from navigation and enhanced situation self-awareness to decision support for spacecraft system design [52], AI is likely to play a major role in long-duration missions to Mars and other locations beyond the Earth-Moon system, especially where significant numbers of crew are involved. These missions would be characterized by crews spending months inside space capsules, where they would be subject to the effects of microgravity, higher levels of radiation and stress. Already, multiple investigations have been mounted where AI is envisaged as a means of mitigating these effects. For instance, in recent years, NASA has begun to explore hibernation as a viable means of keeping crews healthy during long-duration missions. In

2014, NASA partnered with Space Works Enterprises to perform an initial evaluation of a Crew Transfer Vehicle (CTV) where passengers would be placed in a torpor-induced state for the duration of the voyage [58]. The advantages of artificial hibernation extend beyond resource consumptions, aging, and psychology, and include the possibility of improved protection from cosmic radiation. This is based on recent research that relies on early animal models tests that suggest how the effects of radiation could also be reduced during hibernation [59].

Such missions, however, are likely to be heavily reliant on AI for navigation, communication, maintenance and other spaceship operations. A possible solution lies in the form of 'cognitive radio', a technology being investigated by NASA's Glenn Research Centre as a means of increasing the efficiency of space data transmissions. This approach marries advances made in machine learning and cognitive computing to radio communications to handle the heavy volume of communications traffic associated with space missions. It is easy to envision that this combination of AI and space communication could also be applied to long-duration space missions, especially where crews are kept in hibernation. Instead of relying on human controllers, AI-based systems would oversee supplying regular updates to mission controllers and selecting specific radio channels for optimum data transmission.

As Janette C. Briones, the principal investigator in the cognitive communication project, explained:

> The recent development of cognitive technologies is a new thrust in the architecture of communications systems. We envision these technologies will make our communications networks more efficient and resilient for missions exploring the depths of space. By integrating artificial intelligence and cognitive radios into our networks, we will increase the efficiency, autonomy and reliability of space communications systems [60].

Navigation is another area where AI-related research is leading to applications. In 2018, NASA's Frontier Development Lab (FDL) and Intel partnered to develop a system that could assist with navigation on the Moon in the same way that a Global Position Satellite (GPS) assists with navigation on Earth. However, instead of relying on a satellite and tracking software to determine one's location, the system would rely on AI-processed images of the lunar surface. The process of creating this system consisted of using an AI to sort and combine 2.4 million images of the lunar surface, which resulted in the creation of a "virtual Moon" [61]. Based on the team's simulations, this was enough to effectively navigate in lunar environments. The team is hoping to address Mars next, using satellite and rover images of Martian surface to create a "virtual Mars." These and other maps will be incredibly useful when crewed missions are mounted to celestial bodies that do not have a system of satellites yet.

Taken together, autonomous navigation and communication systems could allow for long-duration space missions where crews do not need to be awake for most of the journey. Considering that this will probably be necessary where deep-space missions are concerned (to ensure crew health and reduce the amount of supplies needed), a degree of automation is a must. In addition, AI could play a vital role in ensuring the health and well-being of crews that are kept in waking conditions during long-duration missions. A good example of this is the mobile and autonomous assistance system known as CIMON (Crew Interactive Mobile companiON). This AI assistant, which was developed by the German Aerospace Centre (DLR) in conjunction with Airbus, leverages Watson AI technology from the IBM Cloud and recent developments in robotics to create a voice-controlled artificial intelligence that is fully-autonomous and interactive.

In the summer of 2018, CIMON became the first AI to be deployed to the International Space Station, where it currently aids astronauts with their everyday tasks. Beyond this, CIMON is also a technology demonstrator designed to evaluate the uses of AI in mitigating the stresses of

long-term spaceflight. Its effects on station operations and crew support are currently the subject of an ongoing study [62].

"CIMON is a technology demonstration of what a future AI-based assistant on the International Space Station or on a future, longer-term exploration mission would look like. In the future, an astronaut could ask CIMON to show a procedure for a certain experiment, and CIMON would do that" remarked Marco Trovatello, a spokesman of the European Space Agency's Astronaut Centre in Cologne, Germany [63]. If permanent outposts and/or colonies are created on Solar System bodies, AI-powered assistants and systems are likely to extend to these locations as well. In addition to interacting with crews and monitoring their mental and physical health, AI could be tasked with monitoring a habitat's systems, monitoring vegetable gardens, and sending regular communiques back to Earth.

As humanity's presence in space increases, and missions of greater size and complexity are mounted to locations farther into the Solar System (and possibly beyond), AI will be increasingly relied upon to handle the sheer volumes of data and to assist with complex tasks. The importance of AI in the exploration and colonization of the Solar System was perhaps best summed up by Daniela Girimonte and Dario Izzo (of the European Space Agency's Advanced Concepts Team) in their seminal 2007 study:

> The return of humans to the Moon and a future manned mission to Mars therefore seem to be likely achievements we may witness in the next few decades. At the same time, even more ambitious plans and missions are being conceived by farsighted researchers who dream about the exploration and colonization of even farther planets. In the framework of these more or less concrete future scenarios, the consolidation of artificial intelligence methods in space engineering is certainly an enabling factor [52].

For both a pessimist and an optimist we place space exploration and colonisation on the upward innovation trigger curve expecting the plateau to be reached within the next 5 to 10 years with the optimist slightly higher up the curve.

9. Tomorrow

There is growing, if not irrefutable evidence that activities of human-kind as of approximately 12,000 years ago have increased and continue to increase extinction rates of many species and that we are experiencing the 6th extinction level event (ELE) or the Holocene extinction [64], [65]. The consequences of this latest ELE adds urgency to considering other planets to where mankind can migrate and settle, in part as a potential staging post for further exploration and in part as a fail-safe mechanism should Earth become inhospitable to supporting life as we know it today. This has been exemplified by many scientists and entrepreneurs and two of the most prominent figures are Elon Musk and the late Stephen Hawking. However, with respect to the development of AI, Musk envisages a future where AGI is developed and he has expressed grave concerns stating: "We have to figure out a way to ensure that the advent of digital super intelligence is one which is symbiotic with humanity" [66]. He goes further by saying: "That is the biggest existential crisis that we face and the most pressing one". Likewise Hawking warns: "Someone will design AI that improves and replicates itself. This will be a new form of life that outperforms humans" [67].

Despite the potential, yet theoretical concerns on the development of AGI, Musk and Hawking are joined by Astronomer Royal Sir Martin Rees, physicists Max Tegmark and Michio Kaku among many others who believe that a future for deep space exploration and colonisation to exoplanets *will* involve a human-hybrid avatar and that this may a logical extension of the human species [68-70]. Indeed, based on his assessment of the threat of over-population, depletion of Earth resources and climate change, Hawking has said: "We must continue exploring space in order to improve our knowledge of humanity. We must go beyond our humble planet" [68].

Although AGI is not in our society today, what is in our society and is a subject of considerable debate is defined by the digital after-life industry (DAI). DAI relates to the vast quantity of digital data that are and will continue to be generated at an exponential rate and could be collected and 'reconstituted' after a person dies. This is already an issue for the social platform Facebook, where it is estimated that 1.7 million people per year in the US alone will have passed away and yet their profiles will still be active [71]. The subject of data ownership is an area of intense legal debate and out of scope for this paper, however, what appears to be the case is that the technology platform rather than human person who generated the digital footprint is the owner of the data. This may have some relevance to distant space exploration as a reconstituted digital astronaut or other subject matter expert may constitute a potential future e-crew. Moreover, there is a concurrent philosophical debate on whether a digital reconstitution would have human rights and to demonstrate support for AI in the country, in 2017 a robot named Sophia was given Saudi Arabian citizenship [72]. Another more recent example relates to that of James Dunn, a 24-year-old man who died from the skin condition epidermolysis bullosa and skin cancer. With the help of Pete Trainor from the company Us ai, they were able to develop a chatbot named Bo who was able to replicate aspects of James' thoughts from many conversation James had in 'training' the bot [73]. This is one of several landmark cases which show the potential of the DAI but also seek to call for regulation so that the wishes of both those physically dead and alive are respected and that those who wish to 'live' in a virtual world are accorded the appropriate level of protection and privacy [74], [75].

More immediately, if permanent outposts and/or colonies are created on Solar System bodies, the development of AI systems and assistants is likely to extend to these locations. There are several examples of where AI is being used to make discoveries on space missions [76] where it is proposed that autonomy will be a key technology for the future exploration of our solar system and conditions are not permissible for human habitation and where robotic spacecraft may be even be out of communication with their human mission controllers [77-79].

10. Discussion

In this paper we have briefly considered two aspects of AI that we believe are and will continue to profoundly and positively affect science supporting human healthcare and space exploration. Like many others we can imagine a future of immense benefit and likewise one of catastrophic peril but as scientists we need to be driven by facts as we know them today. The next stage of the digital revolution is happening and is unstoppable, as was the industrial revolution and other paradigm shifts in the advancement of our species throughout the ages. Innovation and adoption of new technology, especially that likely to cause a step change for humanity requires concerted collaboration between optimists and pessimists with critical reality checking, objective communication and an understanding of risk. Let us be bold and work together across our disciplines with dialogue and open collaboration, ensure regulation is in place to understand and plan mitigation strategies and make the very best of a truly life-changing opportunity for our species on Earth and perhaps, in the not too distant future, elsewhere in the Solar System and beyond.

Figure legends

Figure 1. Proposed categorisation of individual populations and their response to and execution of change. Populations may be divided in positive and negative groups, where optimists will likely reside in the positive group.

Figure 2. Proposed superimposition of the Kubler Ross change curve on the hype cycle (purple curve) for A). Individuals who are pessimistic or neither pessimistic nor optimistic (red curve) and B). individuals who are optimistic (green curve). Indicative positions of the two case studies are shown on the hype cycle (yellow star = drug discovery and development for healthcare; red star =

space exploration and colonisation). Differential peak magnitudes shown at the start (A, A^*) and the end (B, B^*) of the cycles and curves are proposed.

References

1. Fagella, D. What is Artificial Intelligence? An Informed Definition, *Emerj*, 2017, retrieved on February 16th 2019, https://emerj.com/ai-glossary-terms/what-is-artificial-intelligence-an-informed-definition/.
2. Turing, A. M. Computing machinery and intelligence, *Mind* 49, 1950, pp. 433-460.
3. Definition of life in English. *English Oxford living dictionaries*, retrieved on February 16th 2019, https://en.oxforddictionaries.com/definition/life.
4. Life. *The free dictionary*, retrieved on February 16th 2019, https://www.thefreedictionary.com/life.
5. Sagan, D, Sagan, S., Margulis, L. Life biology. *Encyclopaedia Brittanica*, retrieved on February 16th 2019, https://www.britannica.com/science/life.
6. Chang, O, Lipson, D. Neural network quine, *arXiv* 1803.05859, 2018.
7. Alasaarela, D. The Rise of Emotionally Intelligent AI, *Machine Learnings* 2017, retrieved on February 15th 2019, https://machinelearnings.co/the-rise-of-emotionally-intelligent-ai-fb9a814a630e.
8. Ghaffarzadeh, K. Mobile Robots and Drones in Material Handling and Logistics 2018-2038, *IDTechEx*, retrieved on February 15th 2019, http://www.idtechex.com/research/reports/mobile-robots-and-drones-in-material-handling-and-logistics-2018-2038-000548.asp.
9. Stanley, K. O., Clune J. Welcoming the Era of Deep Neuroevolution, *Uber Engineering* 2017, retrieved on February 17th 2019, https://eng.uber.com/deep-neuroevolution/.
10. Lehman, J., Chen, J., Clune, J., Stanley, K. O. Safe mutations for deep and recurrent neural networks through output gradients, *arXiv* 1712.06563, 2018.
11. Lehman, J., Chen, J., Clune, J., Stanley, K. O. ES Is More Than Just a Traditional Finite-Difference Approximator, *arXiv* 1712.06568, 2018.
12. Hutson, M. Artificial intelligence can 'evolve' to solve problems, *Science* 2018, doi: 10.1126/science.aas9715.
13. Conti, E., Madhavan, V., Such, F. P., Lehman, J., Stanley, K. O., Clune, J. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. *arXiv* 1712.06560, 2018.
14. Interpreting technology hype, *Gartner*, retrieved February 16th 2019, https://www.gartner.com/en/research/methodologies/gartner-hype-cycle.
15. Sicular, S., Brant, K. Hype Cycle for Artificial Intelligence, *Gartner* 2018, retrieved on February 17th 2018, https://www.gartner.com/doc/3883863/hype-cycle-artificial-intelligence-
16. Amara's law, retrieved on February 17th 2018, https://web.archive.org/web/20180410135130/https://spotlessdata.com/blog/amaras-law.
17. Kubler-Ross, E. *On death and dying*, Routledge, 1969.
18. By, R. Organisational change management: a critical review, *Journal of Change Management* 5, 2005, pp. 369-380.
19. Corr, C. A., Doka, A. J., Kastenbaum, R. Dying and its interpreters: a review of selected literature and some comments on the state of the field. *OMEGA- Journal of Death and Dying* 39, 1999, pp. 239-259.
20. Stroebe, M., Schut, H., Boerner, K. Cautioning health-care professionals: bereaved persons are misguided through the stages of grief. *OMEGA – Journal of Death and Dying* 74, 2017, pp. 455-473.
21. Russell, S. J., Norvid, P. *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, 2003.
22. Hendler, J. Avoiding another AI winter, *Intelligent systems, IEEE* 23, 2008, pp. 2-4.

23. Enwall, T. Why the pursuit of a “killer app” for home robots is fraught with peril, *IEEE Spectrum* 2018 retrieved on February 17th 2019, https://spectrum.ieee.org/automaton/robotics/home-robots/why-the-pursuit-of-a-killer-app-for-home-robots-is-fraught-with-peril.
24. Turck, M. Frontier AI: How far are we from artificial ‘general’ intelligence, really? Retrieved on February 16th 2019, http://mattturck.com/frontierai/.
25. European Parliamentary Research Service. Should we fear artificial intelligence? *European Parliament Think Tank* 2018, retrieved on February 14th 2019, http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_IDA(2018)614547.
26. Fast, E., Horvitz, E. Long-term trends in the public perception of artificial intelligence, In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, Inc*., Menlo Park, CA, 2017, pp. 963-969.
27. Sharo, T., Korn, C. W., Dola, R. J. How unrealistic optimism is maintained in the face of reality, *Nature Neuroscience* 14, 2012, pp. 1475-1479.
28. Hecht, D. The neural basis of optimism and pessimism, *Experimental Neurobiology* 22, 2013, pp. 173-199.
29. Stankevicius, A., Huys, Q. J. M., Kaira, A., Series, P. Optimism as a prior belief about the possibility of future reward, *PLoS Computational Biology* 10, 2014, e1003605.
30. Tufts center for the study of drug development, retrieved on February 17th 2019, https://csdd.tufts.edu/.
31. Hughes, J. P., Rees, S., Kalindjian, S. B, et al. Principles of early drug discovery, *British Journal of Pharmacology* 162, 2011, pp. 1239-1249.
32. Marsden, C. J., Eckersley, S., Hebditch, M. et al. The use of antibodies in small-molecule drug discovery, *Journal of Biological Screening* 19, 2014, pp. 829-838.
33. Perez, H. L., Cardarelli, P. M., Deshpande, S., et al. Antibody-drug conjugates: current status and future directions, *Drug Discovery Today* 19, 2014, pp. 869-881.
34. Valeur, E., Jimonet, P. New modalities, technologies and partnerships in probe and lead generation: enabling a mode-of-action centric paradigm, *Journal of Medicinal Chemistry* 61, 2018, pp. 9004-9029.
35. Valeur, E., Gueret, S. M., Adihou, H., et al. New modalities for challenging targets in drug discovery, *Angewandte Chemie International Edition* 56, 2017, pp. 10294-10323.
36. Monte, A. A., Brocker, C., Nebert, D. W., et al. Improved drug therapy: triangulating phenomics with genomics and metabolomics, *Human Genomics* 8, 2014, 16.
37. Schneider, G., Fechner, U. Computer-based de novo design of drug-like molecules, *Nature Reviews Drug Discovery* 4, 2005, pp. 649-663.
38. Duch, W., Swaminathan, K., Meller, J. Artificial intelligence approaches for rational drug design and discovery, *Current Pharmaceutical Design* 13, 2007, pp. 1497-1508.
39. Olivecrona, M., Blaschke, T., Engkvist, O., Chen, H. Molecular de novo design through deep reinforcement learning, *Journal of Cheminformatics* 9, 2017, 48.
40. Sellwood, M. A., Ahmed, M., Segler, M. H. S., Brown, N. Artificial intelligence in drug discovery, *Future Medicinal Chemistry* 10, 2018, pp. 2025-2028.
41. Segler, M. H. S., Kogej, T., Tyrchan, C., Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks, *ACS Central. Science* 4, 2018, 120-131.
42. Hessler, G., Baringhaus, K.-H. Artificial intelligence in drug design, *Molecules* 23, 2018, 2520.
43. Benhenda, M. ChenGAN challenge for drug discovery: can AI reproduce natural chemical diversity? *arXiv* 1708.08227, 2017.
44. Popova, M., Isayev, O., Tropsha, A. Deep reinforcement for drug design, *Science Advances* 4, 2018, eaap7855.
45. Fleming, N. How artificial intelligence is changing drug discovery, *Nature* 557, 2018, pp. S55-S57.

46. Mak, K.-K., Pichika, M. R. Artificial intelligence in drug development: present status and future prospects, *Drug Discovery Today* 2018, https://doi.org/10.1016/j.drudis.2018.11.014.
47. Pushpakom, S., Iorio, F., Eyers, P. A., et al. Drug repurposing: progress, challenges and recommendations, *Nature Reviews Drug Discovery* 18, 2019, pp. 41-58.
48. Jordan, A. M. Artificial intelligence in drug design – the storm before the calm? *ACS Medicinal Chemistry Letters* 9, 2018, pp. 1150-1152.
49. Lyu, J., Wang, S., Balius, T. E., et al. Ultra-large docking for discovering new chemotypes, *Nature* 566, 2019, pp. 224-229.
50. Topol, E. High-performance medicine: the convergence of human and artificial intelligence, *Nature Medicine* 25, 2019, pp. 44-56.
51. Havelund, K., Lowry, M., Penix, J. Formal analysis of a space craft controller using SPIN, *IEEE Transactions on Software Engineering* 27, 2001, pp. 749-765.
52. Daniela, G., Dario, I. Artificial intelligence for space applications, *Intelligent Computing Everywhere* 2007, pp. 235-253.
53. Weir, N., Fayyad, U. M., Djorgovski, G., Roden, J. The SKICAT system for processing and analysing digital imaging sky surveys, *Publications of the Astronomical Society of the Pacific* 107, 1995, pp. 1243-1254.
54. C. E. Petrillo, Totora, C., Chatterjee, S. et al. Finding strong gravitational lenses in the Kilo Degree Survey with convolutional neural networks, *Monthly Notices of the Royal Astronomical Society* 472, 2017, pp. 1129-1150.
55. Estlin, T. A., Bornstein, B. J., Gaines, D. M., et al. AEGIS automated targeting for the MER Opportunity Rover, *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 2012.
56. Allwood, A. et al. Texture-specific elemental analysis of rocks and soils with PIXL: The Planetary Instrument for X-ray Lithochemistry on Mars 2020, *IEEE Aerospace Conference Proceedings* 2015.
57. Prosser, P., Rebolledo, J. D. AI is kicking space exploration into hyperdrive—here's how, *Singularity Hub* 2018, retrieved on February 17th 2019, https://singularityhub.com/2018/10/07/ais-kicking-space-exploration-into-hyperdrive-heres-how/#sm.000tm2cyt1cylenswso298wt2y6ec.
58. Bradford, J., Schaffer, M., Talk, D. Torpor inducing transfer habitat for human stasis for Mars, *SpaceWorks Enterprises* 2016.
59. Cerri, M., Tinganelli, W., Negrini, M. et al. Hibernation for space travel: Impact on radio protection, *Life Sciences in Space Research* 11, 2016, pp. 1-9.
60. Baird, D. NASA explores artificial intelligence for space communications 2017, Retrieved on February 17th 2019, https://www.nasa.gov/feature/goddard/2017/nasa-explores-artificial-intelligence-for-space-communications.
61. Chung, A., Ludivig, P., Potter, R. W. K., et al. Localization: or the importance of knowing where you are, *Frontier Development Lab* 2018, Handbook, pp. 38-39.
62. Buchheim, J., Alexander, C. Pilot Study with the Crew Interactive MObile companioN (Cimon) (Mobile Companion), *Erasmus Experiment Archive* 2018.
63. Pultarova, T. AI robot CIMON debuts at International Space Station, *Space.com*, retrieved on February 17th 2019, https://www.space.com/42574-ai-robot-cimon-space-station-experiment.html
64. Pimm, S. L, Jenkins, C. N., Abell, R., et al. The biodiversity of species and their rates of extinction, distribution and protection, *Science* 344, 2014, https://doi.org/10.1126/science.1246752
65. Ceballos, G., Ehrlich, P. R., Barnosky, A. D., et al. Accelerated modern human-induced species losses: entering the sixth mass extinction, *Science Advances*, 1(5), 2015, e1400253.

66. Clifford, C. Musk: 'Mark my words – A.I. is far more dangerous than nukes', *CNBC Make It* 2018, retrieved on February 17th 2019, https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html.
67. Galeon, D. Stephen Hawking: "I fear that AI may replace humans altogether", *WIRED* 2017, retrieved on February 17th 2019, https://futurism.com/stephen-hawking-ai-replace-humans/.
68. Tegmark, M. *Life 3.0. Being human in the age of artificial intelligence*, Penguin Press, 2017.
69. Rees, M. *On the future prospects for humanity*, Princeton University Press, 2018.
70. Kaku, M. *The future of humanity. Terraforming Mars, interstellar travel, immortality and our destiny beyond Earth*, Penguin Books, 2018.
71. Evans, C. 1.7 million U.S Facebook users will pass away in 2018. *The Digital Beyond* 2018, retrieved on February 17th 2019, http://www.thedigitalbeyond.com/2018/01/1-7-million-u-s-facebook-users-will-pass-away-in-2018/.
72. Weisberger, M. Lifelike 'Sophia' robot granted citizenship to Saudi Arabia, *Live Science*, retrieved on February 15th 2019, https://www.livescience.com/60815-saudi-arabia-citizen-robot.html.
73. De Quetteville, H. This young man died in April. So how did our writer have a conversation with him last month? *The Telegraph*, retrieved on February 17th 2019, https://www.telegraph.co.uk/technology/2019/01/18/will-digital-soul/.
74. Ohman, C., Floridi, L. The potential economy of death in the age of information: a critical approach to the digital afterlife industry, *Minds and Machines* 27, 2017, pp. 639-662.
75. Ohman, C., Floridi, L. An ethical framework for the digital afterlife industry, *Nature Human Behaviour* 2, 2018, pp. 318-320.
76. Chien, S., Wagstaff, K. L. Robotic space exploration agents, *Science Robotics* 2, 2017, eaan4831.
77. The pivotal role AI plays in the future of space travel. *Ross*, retrieved on February 17th 2019, https://blog.rossintelligence.com/post/ai-space-travel.
78. Campa, R., Szocik, K., Braddock, M. Why space colonisation will be fully automated, *Technological Forecasting and Social Change* 2019; in press.
79. Braddock, M., Campa, R., Szocik, K. Ergonomic constraints for astronauts: challenges and opportunities today and for the future, *Proceedings of the International Conference on Ergonomics and Human Factors 2019, Stratford-Upon-Avon*, 29 April-1 May 2019, 1st Edition.

Studia Humana
Volume 8:1 (2019), pp. 19—48
DOI: 10.2478/sh-2019-0002

De Bello Robotico. An Ethical Assessment of Military Robotics

Riccardo Campa

Jagiellonian University,
Cracow, Poland

e-mail: riccardo.campa@uj.edu.pl

Abstract:
This article provides a detailed description of robotic weapons and unmanned systems currently used by the U.S. Military and its allies, and an ethical assessment of their actual or potential use on the battlefield. Firstly, trough a review of scientific literature, reports, and newspaper articles, a catalogue of ethical problems related to military robotics is compiled. Secondly, possible solutions for these problems are offered, by relying also on analytic tools provided by the new field of roboethics. Finally, the article explores possible future developments of military robotics and present six reasons why a war between humans and automata is unlikely to happen in the 21st century.
Keywords: military robotics, unmanned systems, drones, combat robots, ethical problems, roboethics.

1. Defining Robotic Weapon

Military robotics is a relatively recent phenomenon, and a conventional agreement upon terminology does not yet exist. Therefore, the preliminary praxis in every scientific work, namely to clarify the terms and concepts, is even more necessary in the present context. In US military and political circles[1] the term-concept 'unmanned system' has been introduced to denote systems of weaponry that do not require the presence of human beings where they are located. Such systems are piloted (remote-piloted) *at a distance* by human beings, and even – in the most evolved systems – endowed with greater or lesser autonomy to decide and act. So they are referred to as 'unmanned systems' to distinguish them from 'manned systems,' that is systems without a human operator as distinguished from systems with a human operator. In addition, journalists prefer to use more suggestive expressions such as 'war robot' or 'robot soldier,' even if on closer examination these terms are only used to refer to the more advanced and therefore controversial 'unmanned systems,' that is, those that are of some interest to the press.

In this work we have decided to use the expression 'unmanned system' (UM) as the generic term to refer to any systems of robotic weapon with a military use. We also regard the expressions 'military robots' or 'robot weapons' as being literally equivalent to UM, while the term 'robot soldier' refers only to especially advanced weapons systems, the kind that have some decision-making capabilities, and built for authentic combat.

For a long time, the United States have been compiling and making public a collection of documents with the title *Unmanned Systems Roadmap* that takes stock of the situation on the

features and uses of the military weapons currently available to the army and tracks the future development of these weapon systems over the next twenty-five years. We have roadmaps published on a biennial basis (2005-2030, 2007-2032, 2009-2034, 2011-2036, 2013-2038, 2015-2040, 2017-2042, etc.). The last versions have been called *Unmanned Systems Integrated Roadmap*, because they attempt to integrate the different aspects of the construction and the use of military robots from the point of view of their interoperability. Priority was given to independent accounts and blueprints of the different typologies of military robots that were worked out and then 'added together.'

The *Office of the Secretary of Defense Unmanned Systems Roadmap (2007-2032)* does not give a precise definition of *unmanned systems*, but a definition of an *unmanned vehicle* – the element that constitutes its main component – hints at the meaning. Here is the definition proposed by the document:

> Unmanned Vehicle. A powered vehicle that does not carry a human operator, can be operated autonomously or remotely, can be expendable or recoverable, and can carry a lethal or nonlethal payload. Ballistic or semi-ballistic vehicles, cruise missiles, artillery projectiles, torpedoes, mines, satellites, and unattended sensors (with no form of propulsion) are not considered unmanned vehicles. Unmanned vehicles are the primary component of unmanned systems" [8, p. 1].

So, as well as a positive definition, the vehicle is also given a negative definition, which rules out a whole range of technological products used in war: ballistic vehicles, missiles, artillery projectiles, torpedoes, mines, satellites, static sensors. Positively, these vehicles are ones with their own type of propulsion, that leaves out the human operator, that can act autonomously or be remote controlled, can be reused many times, and can carry a lethal or nonlethal load. They can in fact carry surveillance systems (video cameras, radars, sonars, microphones, etc.) or lethal weapons (cannons, machine guns, missiles, rockets, etc.). The system of a military weapon is defined by the entire vehicle – its form, propulsion system, dimensions, weight, velocity, etc. – and by the load it carries – its electronic brain, its sensors, its weapons, etc. – that together define its belligerent function.

2. Robots of the Sky, the Sea, and the Land

The various editions of the *Unmanned System Integrated Roadmap* offer a large catalogue (albeit incomplete) of robotic weapons systems. Mind that we will not speak about the technical features of every single model, but only of the best known ones. Best known since they have had the honour of media attention precisely because they are ethically controversial in some way or other.

To begin, unmanned systems are divided into three major groups depending on where they are being deployed: in the air, on land, in water. We therefore have unmanned systems equipped for air warfare (UAS – Unmanned Aircraft System), for ground warfare (UGV – Unmanned Ground Vehicle) and for naval warfare (UMV – Unmanned Maritime Vehicle). The latter subdivide in their turn into two categories: Above Water (USV – Unmanned Surface Vehicle) and submarines (UUV – Unmanned Undersea Vehicle). Researchers have renamed UAS as 'flying robots' or 'drones,' a term whose origin is related to the shape of these aircrafts [44], [34], [26].

Also the press has noticed the proliferation of these military robots, as a recent report in an Italian daily attests:

> Bang, a target is hit, no soldiers employed. This is the evolution of defence systems that on-going wars do much to accelerate. Recognition, attack, transportation, tracking and rescuing are tasks more and more frequently given to robots, which paves the way for the automatized warfare prefigured in science fiction movies. Under the generic appellation of Unmanned Systems, these weapons, that function without a human pilot,

> were first in use in aviation and have now been fitted inside motorboats, helicopters and motor vehicles [12].

The article reports that the first 'drone' was used by Israel in the Yom Kippur war, and that sixty years of incessant warfare combined with a cutting edge high tech industry have made Israel the leading nation in the development and production of unmanned weaponry, "surpassing the gutsy US military industry and the land of robotics, Japan." For the sake of precision, it is necessary to recall that "remotely piloted aircraft first appeared during World War I" [48, p. 4]. Remote controlled aircraft were also used by the Americans in the forties, when they tested the effects of the first nuclear bombs. This of course does not intend to diminish Israel's remarkable technological work in the field.

So the article continues:

> During the first Gulf War, in 1991, the Air Force had about a hundred drones; today it is deploying 7000 and keeps churning out new models in response to a demand that knows no limits. This race against the clock is to blame for the high number of accidents: 100 times the number of those involving manned aircrafts according to a study by the Congress. 2009 was the year of the watershed: US aviation trained more pilots in front of a screen with a joystick than in the cockpit holding the control stick. Fewer victims, less expensive to train, but surely more frustrating for Top Gun aspirers.

The article states that there are about forty nations that are developing UM technology. As regards the Italian contribution, it mentions the efforts by Alenia Aeronautica, holder of the patents of SkyX and of SkyY, in addition to taking part in the nEUROn program for the construction of a European unmanned military aircraft and in the Molynx program, the goal of which is the development of a high-altitude robotic twin-motor with up to 30 hours autonomy.

The main goal of the revolution of unmanned vehicles, on land or in the air – the article continues – is that of decreased risk to soldiers. But it is also to contribute, a little like satellites in space, to the complex virtual network of sensors and communications that extend across the stage of operations. Add to this considerations of an economic nature: the take-down of a drone, that flies in any weather, is the equivalent of throwing 45 millions dollars down the drain, if it is a jet fighter 143 millions, naturally not counting the human loss. The US armed force aims for the creation of a fleet of unmanned vehicles equal to a third of the total before 2015. Market valuations predict that turnover in the UM sector may reach 5 billion euros in Europe between 2010 and 2020, and double in the ten years after that and arrive at a total global level of 45 billion euros by 2030.

The article clearly shows one important implication of this new robotic arms race: even if Western nations are at the forefront today, military robots are not the prerogative of these nations, and everything leads one to think that in the future wars will be fought more and more exclusively by machines. More than ever before they will be wars of technology and of industrial systems. Indeed, guerrilla warfare also bets on the potentialities of military robots, so much so that in the Lebanese conflict of 2006 Hezbollah launched 4 drones, the fruit of Iranian technology, on Israeli locations.

Finally, one should keep in mind all the negative and positive repercussions (depending on one's point of view) that the development of military technology has always had on civilians. Even when they are conceived of as systems of weaponry, drones are not limited to military uses. Unpiloted aircrafts are used for the relief work in the case of natural catastrophes and to enforce law and order. For example, the US Coast Guard uses them. New York Air National Guard navy is endowed since 2008 with Predator, an unmanned aircraft nine meters long already used in the war in Yugoslavia. Some models were also in use in Italian skies, on the occasion of the G8 Summit held in Aquila. Remote controlled aircrafts surveyed the crowds spotting turmoil or demonstrators who tried to break into the red zones. "Also the KMax, the unmanned helicopter of Lockheed and

Kaman, is increasingly used to transport gear for the troops, as well as for the transportation of special goods to high altitudes and to intervene in forest fires" [12].

Yet the article by *la Repubblica* mainly focuses on Israel. According to this newspaper, 10 or 15 years from now at least a third of the vehicles in use by the armed forces will consist of UM. Guardium, an armoured vehicle designed by GNius to patrol the borders with Lebanon and Gaza, came into use at the beginning of 2009. It is a small Jeep, similar to a golf cart, fitted with completely automatic command, control and navigation systems. Since last year civilians and the army in the occupied territories have begun using remote-controlled bulldozer convoys to resupply. Rexrobot, a six-wheel vehicle with the carrying capacity of 200 kg of goods to follow the infantry susceptible to receive and execute vocal commands is currently undergoing evaluation in the Israel Defence Forces. Soon will be launched high-velocity unmanned vessels designed by Rafael Industries, with a rigid shell and an inflatable cockpit. The motorboat Protector USV, called Death Shark, is equipped with 4 ultra high definition panoramic cameras (which can capture details 10 miles away) able to shoot in 3D, sonar systems, electro-optical sensors, and remote laser-controlled machine guns able to fixate the target even in rough sea.

What these machines contribute to dangerous demining operations is also fundamental. Many robot models have been designed to explore mined areas and to spot the contrivances. Since mines too evolve – for instance they are now made of synthetic materials that escape the metal detector – the robot's sensory apparatus must similarly evolve to spot these lethal contrivances. For example,

> there are the mine sniffers, that the robotics laboratory of the University of Brescia is currently working on, that use 'artificial noses' pegged to entirely autonomous structures that will recognize the smell of the explosive just like dogs. Researchers in Lausanne have tackled the problem of rough terrain by equipping the mine-seeking robot with mountain bike style wheels fitted with crampons to 'escalate' rocky ground. Today some models even work on solar power [12].

The picture given by this newspaper, even though not always precise and even though it deals exclusively with the Middle East, is detailed and informed enough. Reading the *Roadmap* by the American Department of Defense tells us that the United States pursue the goal of the robotization of the armed forces with a determination no lesser than that shown by Israel. Innumerable prototypes and models are (or have been) produced and used. Here we shall limit ourselves to giving a few examples of each type, in order to give a feel for the technological level that has been reached or that one wants to reach in the future.

2.1. Sky Robots

"An unmanned aircraft system (UAS) is a 'system whose components include the necessary equipment, network, and personnel to control an unmanned aircraft.' In some cases, the UAS includes a launching element" [46, p. 4].

As regards robotized military aircrafts, one model that unquestionably deserves looking into is the MQ-1 Predator, produced by General Atomics Aeronautical Systems, Inc. In use by all three American armed forces, in 2007 120 specimens were delivered, 95 available and 170 commissioned. Since 1995 the Predator has completed missions of reconnaissance and surveillance in Iraq, Bosnia, Kosovo and Afghanistan. In 2001, the US air force fitted Predator with a laser designator to guide ammunition with high precision and enabled it to deploy Hellfire missiles. As a result of these modifications, the machine became multifunctional, that is capable of both combat and reconnaissance. The upgraded version (MQ-1) completed 170,000 flight hours (as of July 2006), of which a good 80% had been in combat. Today the machine has been taken out of service.

Various 'successors' or models developed from the Predator have already been produced by the same company. One of these is the MQ-9 Reaper. In 2009, the inventory of the *Roadmap 2009-*

2034 states that 18 specimens have been delivered and 90 planned. A few years later, the *Roadmap 2013-2038* confirms that 112 vehicles of this type are in service (as of July 1, 2013), and provides a detailed case study of this machine which "illustrates the strategy and actions required, when proper initial lifecycle sustainment planning was not done, to transform the sustainment of unmanned systems from a short-term, rapid-fielding environment to a long-term sustainment environment" [46, pp. 142-144].

The MQ-9 Reaper is a robotic aircraft able to operate at medium altitude, with very high flight autonomy (up to 24 hours). As regards the mission, the priorities have been reversed. This system is primarily a hunter-killer system for critical targets, thanks to electro-optical devices and laser-steered bombs or missiles, with only a secondary role given to the system used in intelligence, reconnaissance and surveillance.

One of the systems adopted by the USAF for high altitude reconnaissance and long flight autonomy is the RQ-4 Global Hawk by Northrop Grumman Corporation (12 machines delivered and 54 planned in 2009, 35 in service in 2013). It is capable of monitoring an area of 40,000 nautical square miles per day, at a maximum altitude of 65,000 feet and with autonomy of up to 32 hours. Surveillance is entrusted to very advanced systems, first tested in 2007: Advanced Signals Intelligence Program (ASIP) and Multi-Platform Radar Technology Insertion Program (MP-RTIP).

Nevertheless, if the principal requirement is to keep the 'spy' flying for many days, without needing to return to base, even for daily refuelling, as is the case for Predator and Global Hawk, the aerostatic robots offer the best performances. Currently the army uses some RAID (Rapid Aerostat Initial Deployment), with a flight autonomy of five days and able to reach an altitude of 1000 feet. This model was used in Afghanistan with decent results. However, much more sophisticated aerostats are being built, such as the JLENS (Joint Land Attack Elevated Netted Sensor), fitted with radar and sensors, able to keep flying at 15 feet for 30 days. Twelve specimens of this model have been planned in 2009. Or the revolutionary PERSIUS of HUAV (Hybrid Unmanned Aircraft Vehicle) typology, manufactured by Lockheed Martin Aeronautics, fitted with sophisticated sensors, capable of flying for three weeks at 20,000 feet without returning to base, and able to move with a hybrid propulsion system.

Other 'flying robots' have shown themselves to be particularly useful to the armed forces because of their small dimension and their easy launch and recovery. In this category we find: small gunships like the Wasp by the AeroVironment, of which almost one thousand specimens have been manufactured; micro-mini aircrafts like the RQ-20 Puma (1137 specimens in service in 2013) or the RQ 11 Raven (7332 specimens in service in 2013); and remote controlled helicopters like the XM 157 Class IV UAS, with 32 specimens provided for the Brigate Combat Team in 2009.

The most futuristic model of robotic aircraft the *Roadmap* mentions is no doubt the X47B by Northrop Grumman Corporation, still at the prototype stage and belonging to the category of Unmanned Combat Aircraft System. Its shape is reminiscent of the interceptor ships of the TV series Battlestar Galactica, and so much so that one might mistake them for an alien spaceship. Only this time the UFO does not contain green men, or men of any other colour. Its captain is the grey matter of the on-board computer. It must be able to take off both from runways and from aircraft carriers, to fly at an altitude of 40,000 feet with 9 hours autonomy, and to carry weapons and bombs of reduced diameter.

Its first ground flight took place took place at Edwards Air Force Base, California, on 4 February 2011. As we read in the Northrop Grumman's website,

> [i]n 2013, these aircraft were used to successfully demonstrate the first *ever* carrier-based launches and recoveries by an autonomous, low-observable relevant unmanned aircraft. The X-47B UCAS is designed to help the Navy explore the future of unmanned carrier aviation. The successful flight test program is setting the stage for the development of a more permanent, carrier-based fleet of unmanned aircraft [20].

Italy as well has a tradition of designing and building robotic aircraft. Feletig [12] only mentions the Sky-X and Sky-Y by Alenia, but the Falco, manufactured by SELEX Galileo and designed by Galileo Avionica, certainly also deserves to be mentioned. It is a small size tactical aircraft designed for reconnaissance and surveillance. Its first flight took place in 2003, but the machine has been officially in service since 2009. Even though the SELEX has not rendered public the name of the user, one knows that five systems (a total of 25 aircrafts and corresponding ground control systems) have been sold to Pakistan. In August 2009 the UM Falco was launched using a Robonic MC2555LLR catapult and has completed the test flight. The first flight by aircrafts fitted with high resolution radar and sensors called PicoSAR (synthetic aperture radar) took place in September the same year. In August 2013, the Selex ES Falco was chosen by United Nations to be deployed in the Democratic Republic of Congo "to monitor the movements of armed groups and protect the civilian population more efficiently" [19]. The Falco flies at 216 km/h and can reach a height of 6500 meters; it is 5.25 meters long and weights 420 kilograms. It is not designed for combat, but a model called 'Falco Evo' fitted with weapons is currently being studied.

2.2. Sea Robots

Unmanned Maritime Systems (UMS) "comprise unmanned maritime vehicles (UMVs), which include both unmanned surface vehicles (USVs) and unmanned undersea vehicles (UUVs), all necessary support components, and the fully integrated sensors and payloads necessary to accomplish the required missions" [46, p. 8].

As regards military robots operating at sea, above or below water, the main mission would seem to be mine hunting. There exists a whole collection of submarines with a shape and propulsion engine similar to those of a torpedo, but fitted with a 'brain' and sensors. The primary task of these machines is the ability to spot mines from among other objects, also taking into account the difficulties specific to marine environments that differ from conditions on land.

Relevant companies are fiercely competing to produce the prototype whose performance will ensure their leadership. Statistics are used to detect the object correctly. Still today it happens that all sorts of objects are mistaken for mines or, worse, that genuine mines are not recognized. We shall not give a lengthy description of the technical features of these machines, but confine ourselves to mention one submarine and one surface vehicle.

Amongst the Unmanned Undersea Vehicles one may take note of the Swordfish (MK 18 Mod 1) by Hydroid LLC, a company that is particularly active in this sector. As for surface vehicles, one example is the MCM (Mine Counter Measures) by Oregon Iron Works, currently in the experimental phase. Surface vehicles for other tasks are also being designed, such as the ASW USV, whose function is revealed by its name: Antisubmarine Warfare Unmanned Surface Vehicle; or the Seafox, an unmanned motorboat specialized in coastal surveillance and patrolling.

2.3. Land Robots

Land robots or, more precisely, Unmanned Ground Systems (UGS) "are a powered physical system with (optionally) no human operator aboard the principal platform, which can act remotely to accomplish assigned tasks. UGS may be mobile or stationary, can be smart learning and self-adaptive, and include all associated supporting components such as operator control units (OCU)" [46, p. 6].

The main mission of land military robots is to clear the ground of mines and explosive devices that are a true nightmare for the Allies' soldiers in Iraq and Afghanistan. Because of their widespread use in this field the MTRS (Man Transportable Robotic System) MK1 and MK2, produced by i-Robot Corp. and by Foster-Miller Inc. respectively, should be mentioned. The *Roadmap 2009-2034* reports that a good 1439 specimens of these machines are already found on the battlefield, but the goal is to roll out 2338 in the coming years. These very useful machines detect and neutralize the explosive devices that military contingents encounter on their path. On the

battlefield 324 MK3s by Northrop Grumman Remotec and 1842 MK4s by Innovative Response Technologies are also in use. These are budget robots that save the lives of a great number of people.

While deminers have been widely used for a long time, the same cannot be said of combat robots (the so called TUGV – Tactical Unmanned Ground Vehicle), that is, of machines with no human operator, that are capable of attacking and killing human beings. Various prototypes are currently studied. One of these is Gladiator, of which six specimens have been produced by Carnegie Mellon University for the Marine Corps. Gladiator is an armed and armoured combat robot, endowed with a collection of sensors and weapons that include: infrared sensors, video camera, rocket launcher and machine guns of type M240 and M249. The vehicle moves on wheels, can be remote controlled by a soldier up to one nautical mile away and is equipped with a system that hides the exhaust gas.

Another machine destined for combat is the Armed Robotic Vehicle (ARV) by BAE Systems, and produced by the US Army. 679 of these have been commissioned in 2009. It weighs 9,3 tons and has been designed to perform two specific tasks. The first is reconnaissance: indeed, the ARV-RSTV model (Reconnaissance Surveillance Targeting Vehicle) is able to scan an area and find, detect and reconnoitre targets with great precision, thanks to its sophisticated on-board sensors. Instead, the ARV-A model is fitted with a range of lethal weapons, among which a medium-calibre cannon, a missile launching system and machine guns. Once the experimental stage is completed, it will be possible to use this model in combat.

However, ground warfare has come to a halt. Among the many reasons one can list the misfortune that happened to Forster-Miller's SWORDS. This is a tiny caterpillar robot carrying a light M249 machine gun. The press and the manufacturer give different accounts, but it would seem that the robotic weapon did not behave as it was supposed to.

On April 11th 2008 *The Register* published a gloomy headline: "US war robots in Iraq 'turned guns' on fleshy comrades." The author tells how the robotic vehicle began to behave unpredictably, stopped obeying orders and spread panic among the soldiers. The tone varies from ironic to apocalyptic: "American troops managed to quell the traitorous would-be droid assassins before the inevitable orgy of mechanized slaughter began… the rogue robots may have been suppressed with help from more trustworthy airborne kill machines, or perhaps prototype electropulse zap bombs" [21].

The news was followed above all by *Popular Mechanics*, which interviewed Kevin Fahey, the US Army program executive officer for ground forces, about this incident. He confirmed it and explained that the robot began to move when it was not supposed to move and did not fire when it was supposed to fire. No human was wounded, but the drill was stopped from precaution. The officer added that "once you've done something that's really bad, it can take 10 or 20 years to try it again" [42].

In reality, in a later article, also published by *Popular Mechanics*, Fahey explained that the SWORDS war robots are still in Iraq, and that they have been neither destroyed nor withdrawn. Cynthia Black, Foster-Miller's spokesperson, also wished to explain that "the whole thing is an urban legend" [42]. Black clarified that it is not a self-driving vehicle. That it can therefore not fire unless told to do so. That the uncommanded movements were due, not so much to the computer going awry, but to a trivial mechanical problem. The robot was put on a 45-degree hill and left to run for two and a half hours, and the motor overheated. When this happens, the engine automatically switches off to avoid breakage. But because it was on a slope the machine started to skid, and gave the impression of autonomous movement. This is the producers' version. Fact is that three SWORDS war robots have really stayed on the battlefield, but placed in fixed positions. Some senior official even wondered if it would not be more practical to put the machine guns on tripods.

So one is given to understand that a hypothetical slowing down of the experimentations is not due to this trivial incident, but to a much more important structural situation, such as the economic crisis that has plagued the United States for the past years and the contextual withdrawal

of US troops from Iraq, expected by the end of August 2010 and completed 10 days ahead of schedule.

Experiments continue in Afghanistan and in laboratories. Also under study is an Unmanned Ground Vehicle, able to spot chemical, biological, radiological and nuclear (CBRN) devices. The iRobot is in fact designing this kind of machine for the US Army.

But the robots on the battlefield can also reveal themselves useful, not only for observing the enemy, unearthing and detonating bombs or fighting. They can also massively assist the wounded during belligerent missions. Many times the wounded cannot be recovered or cured, and therefore they die from blood loss or from the wounds they have incurred, because the place where they find themselves is out of reach or under enemy fire. Here is a machine that can carry out this delicate and dangerous task instead of the stretcher-bearers or of machines operated by humans. Applied Perception Inc. has produced a prototype of Robotic Combat Casualty Extraction and Evacuation. In reality, it is a robot couple. A 'marsupial' vehicle serving as ambulance and connected to a vaguely humanoid machine, with mechanical arms, serving as paramedic. The vehicle is endowed with laser, radar, sensor and systems of navigation that permit it to avoid obstacles and to reach the location of the wounded. In addition the machine is endowed with a telemedicine audio-video system that allows the patient to communicate remotely with a doctor.

The press tells us of other innovative projects that might become a reality in the future for military and civil personnel. All the UGVs mentioned in the *Roadmap* are endowed with wheels, because it does not yet seem that humanoid bipedal models are combat-ready. However, it seems only a matter of time. The performances by the Boston Dynamics 'quadruped' called Big Dog are indeed astounding. In a Fox News footage, Matt Sanchez describes it as follows:

> Using a gasoline engine that emits an eerie lawnmower buzz, BigDog has animal-inspired articulated legs that absorb shock and recycle kinetic energy from one step to the next. Its robot brain, a sophisticated computer, controls locomotion sensors that adapt rapidly to the environment. The entire control system regulates, steers and navigates ground contact. A laser gyroscope keeps BigDog on his metal paws — even when the robot slips, stumbles or is kicked over. Boston Dynamics says BigDog can run as fast as 4 miles per hour, walk slowly, lie down and climb slopes up to 35 degrees. BigDog's heightened sense can also survey the surrounding terrain and become alert to potential danger. All told, the BigDog bears an uncanny resemblance to a living organic animal and not what it really is: A metal exoskeleton moved by a hydraulic actuation system designed to carry over 300 pounds of equipment over ice, sand and rocky mountainsides [27].

This robotic animal cannot fail to attract the attention of the Italian press. Fabrizio Cappella writes in *Neapolis* that "it seems midway between a dog and a giant spider: it has four legs, no head and it walks on broken ground across obstacles: it is called Big Dog and its video has set ablaze the imagination of internet surfers who, for some time now, have fired the wildest comments at the bizarre creature" [6]. The article reveals that this is a project funded by the Pentagon, and that its full name is "Most Advanced Quadruped on the Earth."

Effectively, Big Dog's locomotion is surprisingly natural also on very rough terrain, where it manages to keep its balance in the toughest situations, for example after it has been kicked or after having slipped on ice. The robot moves at about 4 mph and its frame is made of steel: hidden inside, in addition to the petrol engine, are a computer, sensors, video cameras and a global positioning system. It is capable to transport hundreds of kilos of gear and can withstand collision with wheeled vehicles and caterpillars. Its purpose is military; one studies its usefulness to troops in warzones, its ability to carry heavy loads and to transport the wounded. The Pentagon appears to have great faith in the success of the project, given that it has invested 10 million dollars in the prototype. Now in its second version, Big Dog will be further developed and its definitive version

ought even to be able to gallop thanks to the form of its legs that are very similar to those of racing animals.

The comments found online are divided. Some are enthusiastic and others admit to being intimidated and concerned. Here two fundamentally opposed ethical leanings play an important part: on the one hand technophilia (the pride of belonging to the human species, able to build these wonders), on the other technophobia (the refusal to give up pre-industrial or pre-Neolithic lifestyles). What is certain is that this machine, whose locomotion is so similar to that of living beings, does not leave one indifferent.

Another machine that thrills the imagination and stirs up discussion among journalists and readers is a robot called EATR (Energetically Autonomous Tactical Robot), not because of the weapons it carries or because of its locomotive capacities, but because of its system of propulsion and fuel supply. Here it would be appropriate to use the term 'feeding,' which refers as much to machines as to living beings. Effectively the EATR is fuelled much like a human being. Riccardo Meggiato writes in *Wired*:

> Don't be scared if one day, pretty soon, you see a robot among the grazing cows: robots also eat, didn't you know? The EATR, an acronym that does not refer to some exotic train but stands for *Energetically Autonomous Tactical Robot*, is a model that feeds itself: it literally eats plants and converts them into biofuels that it uses to move. The ultimate purpose of this project, still under development, is to create vehicles that cannot only do without classic fuels, but are also able to provide for their own energetic needs. The EATR is the work of *Robotic Technology* in Washington, and its development is funded by *Defence Advanced Research Projects Agency* (aka DARPA). All right, it's the army longing to create autonomous military vehicles, but it is also clear that this kind of technology, once in place, could benefit many different sectors [16].

Wired also reveals some of the technical features of the feeding-propulsion system:

> So how does this cool gizmo work? Oh, it's easy: it forages plants with a mechanical limb and ingests them into a combustion chamber. Once on, it generates heat that warms up reels filled with deionized water, which evaporates. The steam obtained then activates the six pistons of a special engine, which activates an energy generator. This one, finally, is stored in specific batteries and used if needed. To give the system more autonomy, researchers at Robot Technology have developed a range of recovery solutions. For example, if steam escapes from the pistons it is promptly condensed, turned into water and sent back to the combustion chamber. And if there is a shortage of grass to graze, no worries: EATR can happily run on traditional fuels as well, like diesel, petrol, kerosene, and even cooking oil. This, cows can't do [16].

So the robotic system promises a solution to one of the major problems that Unmanned Ground Vehicles encounter: poor autonomy. In order to function on the battlefield, sometimes far from provision lines, it is important not to be constrained by matters of energy. An electric battery may be enough for a vacuum cleaner or a lawnmower, but its performance is unlikely to do for a robotic soldier lost in the Afghan mountains.

Robert Finkelstein, the boss of Robot Technology, guarantees that 68 kilos of plants make the EATR energy autonomous for about 160 km. The vegetarian engine has been construed by *Cyclone Power Technology* from a design by the research centre in Washington. The first experiments predict their integration into a Humvee type military vehicle. Whether it will be mass-produced will depend on the test results, but the producers obviously hope to send the EATR to the battle scene as soon as possible.

Hence Meggiato concludes:

> EATR applications are manifold and go beyond the military. While some have ironically pointed out that they can also be used as weapons to gobble up enemy troops, others view them as tractors ready to work non-stop without refuelling, possibly controlled by another robotic system that does not need human intervention. In the end, whether it is a Terminator or a Winnie the Pooh in high tech wrapping, the future will be vegetarian [16].

3. The Main Functions of the Military Robots

Robots are given the missions that military jargon defines as dull, dirty, or dangerous. In other words, even if some technologies are still not up to replacing man in every relevant task, it does appear rather obvious that the human element is from now on the limiting factor in carrying out certain war missions.

Hard work. Sometimes the battlefield requires work done that the human organism finds it difficult to endure. For example, a human pilot needs a few hours sleep after a long operation; a drone does not. While the longest manned air missions of operation Enduring Freedom lasted around 40 hours, there are now drones that guard some warzones non-stop, remote-controlled by crews on the ground that change every 4 hours. The only limit is aircraft autonomy, but if refuelling can be done in the air then that limit too is removed.

Dirty work. As the *Roadmap 2007-2032* reminds us

> [US] Air Force and Navy used unmanned B-17s and F6Fs, respectively, from 1946 to 1948 to fly into nuclear clouds within minutes after bomb detonation to collect radioactive samples, clearly a dirty mission. Unmanned surface drone boats, early USVs, were also sent into the blast zone during Operation Crossroads to obtain early samples of radioactive water after each of the nuclear blasts. In 1948, the Air Force decided the risk to aircrews was 'manageable' and replaced unmanned aircraft with manned f-84s whose pilots wore 60-pounds lead suits. Some of these pilots subsequently died due to being trapped by their lead suits after crashing or to long-term radiation effects [8].

These incidents persuaded the US military to revert to using robots for dirty work.

Dangerous work. Explosive Ordinance Disposal (EOD) is the primary example of dangerous work entrusted to robots. Improvised contrivances found in the streets and in places where soldiers go constitute some of the major threats in the current military campaigns in Iraq and Afghanistan. Coalition forces in Iraq neutralized over 11,100 Improvised Explosive Devices (IED) between 2003 and 2007. A great percentage of these missions was done by ground robots, and the number of UGVs employed in these tasks has skyrocketed: they were 162 in 2004, 1600 in 2005, over 4000 in 2006, 5800 in 2008.

In order that the performances of military robots meet aspirations, commanders on the field at the head of the different armed forces have been asked to submit a priorities list engineers should focus on. Even though the demands of the ground, air and naval armies differ for obvious reasons, it has become clear that they have four common priorities:

1) Surveillance and reconnaissance;
2) Target identification and designation;
3) Counter mine warfare;
4) Chemical, biological, radiological, nuclear explosive (CBRNE) reconnaissance.

Surveillance and reconnaissance. The main priority has revealed itself to be reconnaissance capacity (electronic and visual). For many army professionals information, qualitative and quantitative, is the key element for operational success and robots are the best candidates to gather this information. The ideal robot is able to exert persistent surveillance (or for long periods) on hostile areas, while

maintaining some degree of ‘covertness.’ Robots are promises that the limits of other systems such as manned vehicles, satellites, submarines and unattended sensors will be overcome.

Target identification and designation. The ability to identify and to locate targets with precision in real time is one of the most urgent necessities on the battle stage. It is necessary to reduce the ‘latency’ and to increase the precision for GPS guided weapons, as well as the ability to operate in high-threat environments without putting warfighters at risk. A quality leap in this sector would improve not only safety, but also be more efficient and efficacious than traditional manned systems.

Counter-Mine Warfare. The most useful yet dangerous mission is that of demining a piece of land or sea. Statistically speaking, since World War II, sea mines have caused more losses of US warships than all other weapons systems combined. The same can be said of landmines and bombs (IED – Improvised Explosive Devices) that are responsible for the majority of losses of the coalition forces in Operation Iraqi Freedom. Commanders regard improving the robot’s capacity to find, tag and destroy these devices as a priority. Henceforth robots appear irreplaceable for this sort of work. They have already saved innumerable lives and, as their technology improves, this ought to reduce casualties still further.

Chemical, Biological, Radiological, Nuclear, Explosive (CBRNE) Reconnaissance. The dirtiest of dirty work is that of spotting CBRNE. Yet this kind of weapon of mass destruction also represents the greatest peril for a nation at war. An attack with nuclear, chemical or biological weapons on foreign land or on troops deployed at the front, would have disastrous consequences not just on the waging of the war, but also on the entire military apparatus, on the economy and on foreign policy broadly speaking. Therefore robots are essential, as much to prevent this kind of attack as to observe and monitor areas that have already been attacked, because of their superior sensorial capacities and because of their greater resistance to chemical, radioactive and microbial agents.

In the *Roadmap 2007-2032* the future goals that constructors and users of robotic weapons in military circles have set themselves are the following:

1) “Improve the effectiveness of COCOM [combatant commander] and coalition unmanned systems through improved integration and Joint Services collaboration” [8]. To this end one expects new designs and experiments on the battlefield with the most promising technologies, accurately testing prototypes prior to their deployment. Reducing the risk in the use of fully developed technologies is also part of the project.

2) “Emphasize commonality to achieve greater interoperability among system controls, communications, data products, and data links on unmanned systems” [8]. Also here the stress is both on security and on safety. On the one hand, it is necessary to improve the ‘common control’ and ‘common interface,’ so that the control systems can easily operate the various kinds of robots. On the other hand, it is important to prevent interceptions, interferences, hijacking, so that the enemy cannot take control of these machines and turn their lethal potential against the army owning it.

3) “Foster the development of policies, standards, and procedures that enable safe and timely operations and the effective integration of manned and unmanned systems” [8]. These goals include:

a) developing, adopting and prescribing commercial and government regulations relative to the design, construction and experimentation of unmanned systems;

b) the coordination between the civil authorities that manage the air, sea and land areas for civil usage (the transport of goods and passengers) and the military authorities in order to prevent collisions between manned and unmanned machines;

c) the development of ever better systems of sensors and control, to give robots the necessary autonomy to avoid collisions with traditional means of transportation.

4) “Implement standardized and protected positive control measures for unmanned systems and their associated armament” [8]. More specifically, one feels the necessity for a standard architecture common to all unmanned systems, armed or not.

5) “Support rapid demonstration and integration of validated combat capabilities in fielded/deployed systems through a more flexible prototyping, test and logistical support process” [8]. More specifically, one intends to develop alternatives to gasoline-powered internal combustion engines, with a particular predilection for high-energy-density power sources (primary and renewable), and if possible common with those of manned systems.

6) “Aggressively control cost by utilizing competition, refining and prioritizing requirements, and increasing interdependencies (networking) among DoD [Department of Defense] systems” [8]. In other words, stimulate both competition among manufacturers and their cooperation, while keeping cost reduction as the primary goal.

New requirements were added to this list, punctually recorded in the updated and integrated *Roadmap 2009-2034*. In particular, one can see that the army insists less on control procedures and security standards, and more on the speedy production of the machines and on their necessary autonomy. This is an important change, which in our opinion reflects the fact that, in recent years, robots came to be viewed as more reliable. So we add the two key points:

7) To maintain the sectors of research and development to increase the level of automatization of the systems of robotic weapons, so that they reach the appropriate level of autonomy, as determined by the combatant for each specific platform.

8) Speed up the transition of robotic weapons systems from the sectors of research and development set up by scientists to the hands of the combatants at the front.

It is therefore considered opportune to maximally stimulate the production and use of ever more sophisticated military robots, because of the army’s ever more enthusiastic reception and implementation of the robots that arrive on the battle stage. Hence moral uncertainties appear to fade away. Besides, operations like demining and the clearance of explosive device in areas either inhabited or in some way traversed by people, as well as the prevention of attack with chemical, biological or nuclear weapons, will hardly raise any ethical objections. What robots do and will go on doing on the field, in time of war and in time of peace, is nothing other than humanitarian work. The same can be said of aid to the wounded. However, it is true that other questions, such as electronic combat and surveillance, could still raise questions of a moral nature. Add to that man’s atavistic fear – symbolically codified in myths, legends and tales – of a rebellion by the creatures against their creator, and one understands that robotic technologies promise to become a main area of applied ethics.

4. Main Objections to the Belligerent Use of Robots

Given that many look upon war as a negation of ethics (sometimes also when it is defensive), and that technological development itself finds firm adversaries on principle, it is not astonishing that the application of robotics to war has stirred up so much discussion [25], [35], [37], [2], [1], [7], [36], [11], [47].

The question however does not engage just pacifists, Luddites, and roboethicists, but also military professionals and engineers. The development of this kind of operation does indeed promise to solve many problems, but it is not without its pitfalls. The debate is therefore more necessary than ever. Here we shall outline the moral objections to the use of military robotics that we have found most cogent, and, in a second part, we shall evaluate them both from a technical and ethical point of view.

4.1. Noal Sharkey’s Plea

A plea by the Royal United Services Institute (RUSI) that denounces the dangers of a robotic arms race and the risk that it would imply for all humanity has caused a particular stir in the media. The plea has made headlines because it is written by experts in the new technologies and, moreover, for so prestigious an institution as the RUSI. Those who are informed about military matters know well that the RUSI is not some hangout of pacifists or Luddites.

This point of view has found one of its more notable spokespersons in Noel Sharkey, professor of Computer Science at the University of Sheffield. According to him,

> the trouble is that we can't really put the genie back in the bottle. Once the new weapons are out there, they will be fairly easy to copy. How long is it going to be before the terrorists get on in the act? […] With the current prices of robot construction falling dramatically and the availability of ready-made components for the amateur market, it wouldn't require a lot of skills to make autonomous robot weapons [39].

The first argument that the anti-robot front puts forward is therefore the possibility that the enemy could use these creatures against us. Strictly speaking, this is a prudential argument rather than an ethical one. Indeed, it is about our own good, rather than the good of other fellow humans. There is a fear that our own drive for hegemony can turn against us. Western nations are apparently investing huge amounts of money in the construction of these war machines (4 billion dollars in 2010 and a total expense of 24 billion dollars in the case of the United States), but once they fall into enemy hands they are easy to copy. At what point will Islamic fundamentalists or other enemies of the West no longer need kamikaze and suicide bombers, but will be able to direct remote controlled drones with lethal charges against preselected targets? Sharkey has been interested in this problem for a long time, and also worked as an advisor to the BBC during the broadcast of the television series *Robot Wars.*

Maruccia observes that "the professor does not give much detail as to this presumed facility to build, but he does assure us that a drone equipped with an autopilot guided by Sat Nav currently carries the modest price tag of 250 dollars" [15]. He probably refers to mini drones, given that a Predator costs around 4 million dollars, but we can certainly bet that the cost of these technologies will fall substantially. In addition, it is true that mafias and terrorist groups sometimes dispose of large sums of money and that, for the sum that a Nation spends on the purchase of a supersonic jet plane, one can buy 30 Predators.

The second ethical problem that Sharkey brings up is the drones' limited capacity to discern, that is, the possibility of error: because of the 'relative blindness' of the machines currently in use it is not possible to guarantee the discrimination between combatants and innocents or a proportional use of force as required by War legislation:

> Allowing them to make decisions about who to kill would fall foul of the fundamental ethical precepts of a just war under *jus in bello* as enshrined in the Geneva and Hague conventions and the various protocols set up to protect civilians, wounded soldiers, the sick, the mentally ill and captives. There are no visual or sensing systems up to that challenge [30, p. 87].

In an article appeared a few months later on *Science*, Sharkey clarifies that "even with a definition [of a noncombatant], sensing systems are inadequate for the discrimination challenge, particularly in urban insurgency warfare" [31].

Here the misgiving is mainly ethical, because it concerns others' safety. But let us add that the error could also consist in killing allied soldiers. The so-called friendly fire. Because of this, Sharkey solicits a serious international debate, one which also takes hypotheses of a moratorium into consideration: "With prices falling and technology becoming easier, we may soon see a robot arms race that will be difficult to stop. It is imperative that we create international legislation and a code of ethics for autonomous robots at war before it is too late" [29]. In other words, the international community should evaluate the risks of these novel weapons *now*, rather than sit around and wait while they sneak their way into common use.

4.2. Robotic Wars as War Crimes Without Criminals?

The question of the possibility of errors is raised also by Andrew Brown in a blog related to *The Guardian*. However, he lays the stress above all on the matter of relieving oneself from the burden of responsibility. Reflecting on the concept of hostile artificial intelligence, Brown warns that the robot has a particular status that it is hard to define: it is not yet a sentient being capable of moral discernment, but neither is it a mere object controlled by man: it has a goal and it pursues it, even though it does not know that. By the way, this is true also for the so-called smart bombs, that follow heat or satellite signals: "The missile, a thing that is both dead but none the less animated by a hostile purpose, violates some primitive expectations of the way the world works. That's one reason it seems so frightening" [3].

Brown raises the problem of the moral status of the robot, of the people that are constructing it, that give the order to use it, and that use it. He rejects the idea of those he calls "the protagonists of extreme artificial intelligence," for whom the robot is considered on a par with humans once its behaviour becomes indistinguishable from that of a human (that is, that it passes the famous Turing test). He therefore proposes a further ethical problem linked not so much to the blindness as to the possible lunacy of the robotic soldier. He asks

> what would happen to a robot which acted against its programmers' intentions: if it started to shoot everyone less than four feet high, or offer sweets to anything armed with an RPG?[2] The answer is obvious. It would be either reprogrammed or destroyed. A human, on the other hand, would be tried, because a human could be blamed – or praised for what he had done.

According to Brown, an entirely unprecedented problem arises in the presence of hostile artificial intelligence: we could have war crimes without the possibility of identifying for certain the war criminals.

4.3. Trivialization and Multiplication of Armed Conflicts

Also Peter W. Singer has dealt with this question in a lengthy and detailed article that appeared in *The Wilson Quarterly* [32]. Singer begins by describing the death of a soldier, one much appreciated by his fellow soldiers and by his commander for his courage, tenacity and ability. He had saved many lives but, during a demining operation, the device that he was trying to deactivate exploded, killing him. His comrades in arms picked up his remains and carried them away from the scene by helicopter. When writing their report, the commander lavished words of praise and gratitude for the soldier that offered his life, but said that, at least, there was one thing he was relieved about: "When a robot dies, you don't need to write to its mother."

The death of PackBot cost US taxpayers 150,000 dollars. It will be replaced with few tears shed by its clone. Or by a more advanced model.

Singer starts out with this example to argue that robotic war opens up new sociological, psychological, ethical, legal, and political scenarios. A novelty comparable to that offered by World War I, the first major conflict after the industrial revolution. Drawing inspiration from science fiction writers of the time (H. G. Wells, A. A. Milne. Arthur Conan Doyle, Jules Verne, etc.), farsighted politicians like Winston Churchill and engineers tried hard to put previously unseen 'steel monsters' on the battlefield: armed tanks, aeroplanes and submarines. This brought war to a level it had never reached before. The biggest novelty was that the new weapons (machine guns, gas, armoured tanks, etc.) made a carnage of any attempt to move the front just a few meters, while planes and zeppelins managed to bring the war from the front to inhabited cities and unarmed civilians, and submarines came to threaten passenger ships and unarmed freighters. It radically altered the way in which war was fought, and not just as regards the strictly technical, but also the human.

The same is happening now with robotic arms. Singer, even though he underlines the obvious positive aspects of these weapons for whoever has them, that is, that they spare human lives in one's own faction, he brings up another question for ethical discussion, that someone has called 'the videogame effect.' Those who fight with robotic means are very far from the battlefield and do not always feel as if they were killing living and sentient beings. We could refer to this problem with an expression: 'trivialization of war.'

The testimonies that Singer collects give a fairly clear idea of the new psychological dimension the fighter finds himself in. While the Predator's sensors spot the enemy on the mountains of Afghanistan and attack him with lethal weapons, the human pilot is 7500 miles away in a military base in Nevada. The experience is that of a psychological disconnection between being 'at war' and leading a normal working life. A pilot of Predator describes the sensation as follows: "You see Americans killed in front of your eyes and then have to go to a PTA [Parent Teacher Association] meeting." Says another: "You are going to war for 12 hours, shooting weapons at targets, directing kills on enemy combatants, and then you get in the car, drive home, and within 20 minutes you are sitting at the dinner table talking to your kids about their homework" [32].

Another interesting question that Singer raises concerns control, human presence in decision-making. This is the EURON Codex's first point: "Safety. We should provide for systems for the control of robots' autonomy. Operators should be able to limit robots' autonomy when the correct robot's behaviour is not guaranteed" [40, p. 617]. Or, to say it like Eliot Cohen – an expert on military questions who has worked in the State Department under the administration of George W. Bush – "people will always want a human in the loop." Although we may want this, it is time to ask if this is technically possible, if it will not lead to rather paradoxical situations.

In fact, as the number and quality of robotic arms improve, humans will get expelled from the 'loop' little by little. This process was visible already at the time when electronic weapons emerged (radar, radio, sonar, etc.) in the first half of the 20th century [1], and is becoming ever more visible today. Let's begin with an example. During the Gulf War, the captain and radar navigator Doug Fries describes bombing operations as follows: "The navigation computer has opened the aircraft hold door and unhooked the bombs into the dark." Of course other human beings programmed the machines initially, but then one allowed the computer to take over on the battlefield, giving the pilots a merely auxiliary role.

The most tragic event in connection with this kind of procedure took place also in the Persian Gulf in 1988: the case of Iran Air Flight 655. In the eighties US naval ships had been endowed with the computerized defence system Aegis that had four different modalities of action. Among these were the 'semi-automatic' modality, which gave humans the possibility to decide if and what to fire at, and the 'casualty' modality designed to run the ship and defend it if all the men on board were dead. On July 3rd 1988, the USS Vincennes, renamed Robo-cruiser for the Aegis system and because of the captain's aggressive reputation, detected the presence of an aircraft and identified it as an Iranian F-14, and therefore signalled it as an 'assumed enemy.' Although Aegis was set in 'semi-automatic' mode, that is, with the machine given minimum decisional autonomy, none of the eighteen marines and officers of the command wanted to take the responsibility of contradicting the computer. Hence they followed its advice and authorized fire. The missile destroyed an innocent passenger plane with 290 passengers on board, among which 66 were children.

Let us therefore make a list of the errors made:

a) The Aegis is designed to oppose the action of Soviet bombers in the north Atlantic in war time and acted according to these directives, and yet it found itself beneath a sky full of civilian planes in peace time;

b) His great trust in computers lead the commander to drop a security procedure that envisaged asking higher officials on other war ships for permission;

c) once again, deep faith in computer wisdom induced the captain and his collaborators to blindly listen to the advice of the machine, despite the improbable nature of an Iranian attack.

Similar errors have occurred with other robotic or automatic weapon systems. During 2003 the invasion of Iraq, a battalion of Patriot missiles took down two allied aircrafts upon having mistakenly classified them as 'Iraqi rockets.'

Here then is what the situation looks like, beyond the problems. In theory, humans are still in the loop, part of the decision-making, but the truth is that decisions have to be made in seconds, between the computer signal and the possibility of one's own death, and therefore no one feels up to using what now boils down to a 'veto power.' One always allows the robotic weapon to fire and hopes that it will strike the enemy and not unarmed civilians or allies. When acting under such psychological stress, it is as if humans had no role to play.

This situation is summed up in what we could name the 'paradox of controlled autonomy.' Many have become aware of the problem, among these the psychologist and expert on artificial intelligence Robert Epstein:

> The irony is that the military will want it [a robot] to be able to learn, react, et cetera, in order for it to do its mission well. But they won't want it to be too creative, just like with soldiers. But once you reach a space where it is *really* capable, how do you limit them? To be honest, I don't think we can [33].

In other words, one first constructs a machine able to do things humans cannot, and then one still expects that humans would have the last word about what the machine ought to do. This is paradoxical.

The result is that, when releasing thousands of robots on the battlefield, one continuously feels the need to introduce exceptions to the general rule that wants humans to have the last say in all decisions. Let us look at it in more detail, reasoning in terms of degrees of autonomy, and not just in a digital one/zero perspective.

First exception. Just as an official has authority over a certain number of human soldiers, one imagines that an operator could supervise a certain number of robotic soldiers. The problem is that the number of robots that a human being can control is directly proportional to the individual robot's degree of autonomy. To understand the problem, let us imagine that we are playing five videogames at the same time. A Pentagon report stresses that "even if the gunship commander is aware of the position of all his units, combat is something so fluid and rapid that it is very hard to control." In other words, if we really want them to fight, and if we cannot assign one commander to every robot, we have to give them the possibility to respond autonomously to enemy fire.

Second exception. No reminder is necessary that the enemy is as sentient and uses electronic arms just as much as we do. As early as the Tsushima battle of 1905, Russians and Japanese used radio waves to spot their mutual presence or to interfere with the communication between battleships [10, pp. 66-74]. If the robotic soldier cannot fire unless a remote operator (a human soldier) authorizes it, then it will be enough to obstruct the communication to render the machines harmless and leave them at the mercy of the enemy. In other words, it makes sense then to set up a plan B in the case communications are cut off which envisages the possibility of robot decisional autonomy. In this case they will be able on their own to defend themselves against threats, hit the enemy and return to the base. We can only hope that they make no mistake.

Third exception. Even if every robotic weapon has its own operator, even if the communication is not broken, even if the enemy does not operate at digital speed, there are situations in combat in which humans cannot react fast enough to neutralize a threat. If a projectile is fired at a robot, it takes a human some time to notice (due to the time of propagation of sound waves, to brain reaction time, to momentary inhibition provoked by noise or fear, etc.), while a robot is at once able to spot the conflagration source and frame it as the target of a laser ray. If one can point a laser at someone who fires, then in the same way one can fire a lethal projectile. That is, if one is working in auto-mode, without waiting for a human operator to give the green light, then one could shoot down anyone firing before he had the time to put away his weapon and hide or run away. It is a very strong argument that soldiers on the field are quick to point out. Which human

being would risk life and limb, with a very high probability of instant death, in order to kill a machine? Giving the robot enough autonomy to return the fire would totally change war and guerrilla warfare. It would make armed insurgence pointless, because this one is linked to a need for revenge on the occupying forces. Among other things, introducing this exception could seem attractive not just to soldiers, but also to public opinion, which looks rather favourably at the asymmetry between attacking and responding to an attack (even in a 'superhuman' way). The robots do not aggress humans, but eliminates them if they become aggressive and dangerous.

One is also considering the hypothesis of taking exception to the general rule of control, in partial terms, that is, enabling the robot to fire, but only in order to strike machines and not human beings (other robots, armoured tanks, jeeps, etc.). In this case, a robot could block the enemy by targeting wheels or caterpillars. However, in so doing, the robot would not shelter from enemy fire, given that the human operators would or should survive. And it would not shelter fellow soldiers, given that survivors would keep their ability to fire and kill. The dilemma does therefore not go away, and the idea generally speaking of an exception remains a sensible one.

The problem is that, by multiplying exceptions, one risks giving full freedom to the machines. As robotic weapons become more and more reliable, commit fewer and fewer errors, they will get to so high a degree of reliability and, in combination with the technical impossibility for man to replace the machine, we will reach a point of no return. Let us not forget that, indeed, humans too make mistakes. Military history is rich with episodes of friendly fire being more homicidal than enemy fire. Humans are not less dangerous than computers or robots. Even in the presence of errors, it will be enough that statistics weigh the balance in favour of the computer or the robot, to completely remove humans from both battlefield and decision-making.

What might happen in the future, starting with these observations and looking at the technological trends, it is the emergence of yet another ethical problem: the increase of belligerent conflicts. This at least is the opinion of Lawrence J. Korb, an ex marine officer, the author of some twenty books, who has served also as assisting secretary of defence during the Reagan administration. Korb is a great supporter of robotic weapon systems because these save human lives. However, he is persuaded that this is precisely why technological development will make it ever easier psychologically to decide to go to war. There are two factors that push in this direction, and both are the effect of the automation of the armed forces: a) The growing disconnection between the military apparatus and civil society; b) The perverse voyeurism to which emerging technologies give rise.

As Singer reminds us,

> Immanuel Kant's *Perpetual Peace* (1795) first expressed the idea that democracies are superior to all other forms of government because they are inherently more peaceful and less aggressive. This 'democratic peace' argument (cited by presidents across the partisan spectrum from Bill Clinton to George W. Bush) is founded on the belief that democracies have a built-in connection between their foreign policy and domestic politics that other systems of government lack. When the people share a voice in any decision, including whether to go to war, they are supposed to choose more wisely than an unchecked king or potentate" [32].

In other words, since we know that war can bring both victory and glory, or death and despair, and since it directly affects citizens and all their loved ones, democracies strongly pressurize their leaders and urge them to caution and to responsibility rather than to irresponsible adventures. Indeed, glory is mostly to the benefit of the leader, while the loss of loved ones befalls the ordinary citizen. Not forgetting that, in past wars, citizens who had stayed at home, even if they had no friends or relatives at the front, had to face rationing of certain products of consumption (food, clothing, gas) or pay a war tax to sustain the war effort.

But what happens if one sends mercenaries and robots to war instead of citizens, and that one has to put up with neither taxes nor rationing? There will be a general disinterest in the war.

One is reminded of it only an instant at the airport when one's toothpaste is confiscated because it exceeds the 100 ml limit. In any case, the influence of public opinion in democratic nations is more theoretical than a reality. The United States of America have fought on many fronts in the last half century, from Korea to Vietnam, from the Persian Gulf to Yugoslavia, from Afghanistan to Iraq, not counting all the minor interventions in Latin American nations. However, the last formal declaration of war goes back to 1941. Italy as well has circumvented the constitutional obstacle that only allows for defensive war and classified foreign interventions (the Gulf War, the attack on Yugoslavia, the invasions of Afghanistan and of Iraq, the intervention in Lebanon, etc.) as 'international police operations' or as 'humanitarian interventions.'

The argument put forward by Korb, Singer, and other experts in robotic weaponry is therefore the following: if 21st century wars no longer require the approval of Congress, if there is no rationing, if no special taxes are imposed, and last but not least machines are made to fight instead of humans, then political leaders will be ever more at liberty and have ever better reasons to opt for military interventions.

To give just one example, faced with the massacres in some of the African nations that we have recently observed (think of the ethnic cleansing done in Rwanda, with children and grown ups actually beaten to death with machetes), Western nations have felt impotent. It could have been politically risky to send troops (perhaps via conscription) into such tough conditions, even with the good intention to save children and innocents. Massive losses would lead to electoral defeat of those politicians taking that decision. But if we had the 'robotic weapons of the future,' the decision might have been another. Predators and Reapers, controlled from Nevada or a European base, could have massacred the Rwandese irregular military bands, and saved the lives of many unarmed civilians, without jeopardising the lives of compatriot soldiers. Therefore this is an attractive argument that it will be ever harder to resist, both for the government and for public opinion.

The second issue Korb raises is that of technological voyeurism. Today Predators see the enemy and kills it. They do exactly what humans used to do at the front. The difference is that human soldiers stored these cruel images inside their brains, that is, in hardware that does not allow file sharing (for the moment being at least). They could tell of what they had seen in their war diaries, or on the radio, or on television. But no one else could see it like they had. Today thousands of movie clips made by drones end online, especially on Youtube, visible, downloadable and distributable by anyone. The military calls these video clips 'war porn' because they show all the cruelty of war with no censorship. People – also because in fiction films they are constantly exposed to violence and phenomena such as spurting blood and exploding brains – are not particularly impressed with death on live. As an example, Singer refers to a video in which a Predator strikes a group of insurgents, having their bodies bounce into the air, while one hears the tune of a pop song by Sugar Ray with the title "I Just Want To Fly." This way war is almost transformed into a sport event, a show, in which the audience is ethically numb, cruel, hungry for revenge, and wants entertainment, and feels none of the compassion that one would expect.

This also happens because the US authorities filter the images and only let through those that serve propaganda. The images that show American soldiers hit, mutilated or killed by the enemy are censored. It would be hard to watch a friend, a son or just someone one knows bounce in the air and pass from website to website, in order to satisfy this kind of pornographic voyeurism. Relatives or friends would have the clip removed. Besides, psychologically, it could have all kinds of effects and unpredictable responses: on the one hand it could increase the desire for revenge, on the other hand it might convince public opinion that war is a pointless bloodshed (that of friends or of the enemy).

War reduced to a videogame, with appropriate filters, could act favourably on public opinion and the ruling classes. Thus, paradoxically, the development of robotic weapons, through decreasing the cost of war in human lives and stress, could in the future increase the number of conflicts as whole, and so increase the level of existential risk for all humanity.

But this is not the only ethical problem. In Singer's words, "such wars without costs could even undermine the morality of 'good' wars" [32]. A nation's decision to enter war, in order to

assist another country that has been aggressed and is close to succumbing, is a moral act especially because that nation is not directly threatened. The moral act lies in the disinterested risk that it takes to lose lives and money. The moral act lies in the collective choice and in the price paid. But if both choice and losses vanished, where is the moral act?

> Even if the nation sending in robots in a just war, such as stopping genocide, war without risk or sacrifice becomes merely an act of somewhat selfish charity […] The only message of a 'moral character' a nation transmits is that it alone gets the right to stop bad things, but only at the time and place of its choosing, and most important, only if the costs are low enough [32].

5. Analyses and Propositions

We have included enough arguments of the ethical kind, for or against the use of robotic weapons. We shall now examine them in the light of the principles and the ethical codes that have been elaborated by roboethicists in recent years [41], [13], [4], [38] and in particular the already mentioned *EURON Roboethics Roadmap* [40].

5.1. The Impracticability of the Moratorium

One has, first of all, proposed to bring the robotic arms race to a halt via a moratorium or a ban. Professor Noel Sharkey has formulated the question in precautionary terms, saying in essence that we should hesitate to produce these weapons because they might fall into enemy hands. But this assumes, as its starting point, that only the West is implicated in the manufacturing of these weapons and that hence it is enough to address the editors of the *Roadmap* and a few others to forestall the peril. In reality many nations have for decades been working on robotic weapons systems. As we have seen, drones have already been used, in the 1940s, by the Americans and in the Yom Kippur War by the Israelites, and in addition the Hezbollah in Lebanon and the Pakistanis also have them. It is hard to believe that the Russians and the Chinese have renounced them. It is necessary to understand that there is more than one player and, consequently, no matter how sensible the arguments of the robo-sceptics are, we find ourselves in a classical strategic dilemma, which makes it impossible to make a just choice, for structural reasons that are independent of any single will.

The model is in fact similar to the so-called 'prisoner's dilemma,' an indispensable case-study in every textbook of practical ethics, as well as the basic problem in game theory, which demonstrates how two people (or parties, armies, nations, etc.) might not cooperate, even when cooperation would be in the interest of both.[3] One example of the prisoner's dilemma is the following. Two criminals are arrested by the police. The police does not have sufficient evidence to incriminate them, so it separates the prisoners and visits both of them and gives them the same deal: if the one witnesses in favour of the incrimination of the other (that is, if he defects) and the other remains silent (that is, cooperates), then the accuser is freed and the silent accomplice gets ten years. If both remain silent, both prisoners get just six months in jail for a minor offence. If both betray the other, each is condemned to five years incarceration. Each prisoner must choose whether to betray her/his accomplice or keep quiet. Each is assured that the other prisoner will not be informed that (s)he has been betrayed before the end of the investigation. How should the prisoners act?

Various philosophers and mathematicians have tackled this problem, among whom John Nash, who formulated a solution known as 'Nash's Equilibrium.' One generally agrees on the most likely result of the negotiation. If one assumes that all each player wants is to minimize his own time in jail, it follows that the prisoner's dilemma does not form a zero-sum game, in which each player can either cooperate with the other player, or betray her/him. The only equilibrium in this game is a so-called 'Pareto suboptimal' solution, in which the rational choice induces the two

players to defect, and get five years, even if the gain for each player would be superior if they cooperated (for just six months).

This dilemma had much success, partially because it was formulated during the Cold War and appeared as a perfect description of the arms race between the USA and the USSR (the two prisoners). It was in the interest of both to stop the race, but mutual lack of confidence impeded cooperation. Nothing has changed much with the robotic arms race, with the difference that now the prisoners are not just two, but many. This renders the solution to the problem at the mathematical level even more complicated.

This is not to say that it would be naïve or useless to state the problem, but simply that it would be naïve to believe that there is an easy solution for it, or that the ethical problem is just a dilemma with a binary choice. To think that one can stop the robotic arms race with a proclamation is like imagining that shouting: "Crimes must cease!" on the rooftops will defeat crime. Crimes can be defeated only if one removes the causes that generate them and at the same time makes sure that the victims and the suspected criminals are not denied their rights. The same goes for robotic weapons. As long as there are wars, nations involved will always want to have the most powerful and sophisticated weapons. So, if these frighten, then one needs to envisage creating a balance of geopolitical forces that makes resorting to war rare and inconvenient. Crying wolf is not enough. We need (and this is a lot harder) to find the lair and tame it.

If convincing a nation to renounce making robotic weapons may seem all but impossible (the reply will be: "Go convince Russia and China, and then we'll talk about it"), the idea however of opening up a debate to regulate its use, also in wartime, is not futile. The same goes for chemical, bacteriological and nuclear weapons. To conclude, one may accept the idea of not using them, but not the idea of not having them.

The goal of 'owning with inhibited use' is perfectly in line with the principles of rational ethics. And it is also compatible with Immanuel Kant's approach to ethics, as well as with some of the principles of ancient traditional morality – Eastern and Western. Effectively Kant's meta-norm known as the 'categorical imperative' can be formulated as follows: "Act in such a way that the maxim of your (subjective) action could become a universal (objective) law." In spirit, if not in letter, it comes close to the principle of reciprocity of Confucian tradition (embedded also in the Gospels): "Do not do to others what you would not want done to you." Applying the categorical imperative (or the principle of reciprocity) to one's own actions, anyone can see if these are moral or not. Thus one could ask a thief: "Would you want burglary to become a universal law, that is, that everybody would steal instead of doing honest work?" It is obvious that *rationally* the thief would have to give a negative reply, because if everybody stole there would be nothing to steal. If rendered universal, the immoral act becomes impossible.

Of course in war the principle of reciprocity has been often violated. In addition, also at the theoretical level, everybody does not accept the idea that ethics must have a rational foundation or be founded on an egalitarian principle such as the one just outlined. Those who view themselves as 'the elect people' or 'a superior race' or 'a nation with a manifest destiny' could give themselves rights and prerogatives that they do not concede to others. But what we want to stress here is that, contrary to what one might think, an egalitarian approach to ethics does not at all rule out belligerent action. The categorical imperative is meaningful also in the context of war and is compatible also with military operations. We will give just one example. We can kill our enemies in a gunfight, with conventional weapons, also because we have accepted the possibility of dying in such a context. However at the same time we can refuse to pluck out our enemies' eyes, because we would never want this to become a universal law and that our own eyes were plucked out, should we become prisoners. In the end, the purpose of a rational approach to ethics is that of creating conventions and rules that are widely shared, also in situations of lethal conflict. And history demonstrates that this is not a chimerical approach. Even during World War II, which, by virtue of its use of devastating weapons and the total number of casualties, has been the most bloodthirsty conflict in human history, none of the fighting powers – however radical the ideological

confrontation – violated certain conventions and rules that had been agreed upon: for instance the prohibition to use nerve gas on the battlefield.

To sum up, because of the prisoner's dilemma, it makes little sense to require that nations forgo robotic weapons, especially now that we find ourselves in a phase of history with many and widespread conflicts, but that because of Kant's principle of the categorical imperative, as shown by various historical cases, it becomes possible (and also cautious) to arrive at an international convention that regulates the use of these weapons.

5.2. Pragmatism As a Remedy for Undesired Effects

The second major issue Sharkey, Brown, Singer and many others raise has to do with robot errors, to their hypothetical going awry, to the problem of defining the responsibility in the case of slaughter of innocents (as in the emblematic case of the Iranian Airbus in 1988). This is a serious problem with no easy solution, which has occupied both commissions and magistrates. If it is not possible to punish the robot, then it is clear that the responsibilities can be shared (according to the case) among designers, makers and users, as happens with other technological objects.

However let us make one thing clear: the hypothetical elimination of electronic equipment and automatic systems from airplanes, warships and battle tanks does not at all shelter us from possible errors. Human beings are also prone to errors and, worse, deliberate cruelty. When one repeats like a mantra that "the control must remain in human hands," in order to reassure public opinion, this one should ask itself which human hands will indeed control the weapons. Robots may kill civilians by mistake, which indeed is awful, but let us not forget that humans have systematically and deliberately killed civilians out of revenge or cruelty. Think only of indiscriminate bombing of cities in order to sap enemy resistance.

The robot soldier might mistakenly point his weapon at a civilian or kill an enemy that has already surrendered, but the human soldier is capable of worse. He has tortured prisoners, humiliated, mutilated and killed them for the sheer pleasure of it. We can mention the Turks who impaled prisoners, or the Phoenicians who during the Third Punic war mutilated Romans on the walls of Carthage and threw their remains into the air. Or had them crushed by elephants or by the keels of their ships. But without going back this far, it is enough to think of the tortures some US soldiers inflicted on Iraqi prisoners.

Finally, it may be fruitful to discuss the possibility that robots change (blindly) into potential assassins, but we do not think that these problems could be resolved by simply handing control over to humans. Humans are not angels. They could commit atrocities that are much worse than machine errors. Add to this the fact that technology keeps improving, while humans evolve much more slowly, and the argument from error might be overcome in a couple of decades. In other words, one should not think of control as the negation of all autonomy, but rather as the capacity to stop the machine from functioning should the situation degenerate dramatically.

To put it in even clearer terms, the ethical and cautionary problem, once one has adopted a pragmatic perspective, is not resolved by imposing human control as a matter of principle, but by continuous assessment (and so the old procedure of trial and error), which is the procedure that would offer the best results, that is, to achieve our ends with the fewest casualties, both friendly and enemy. This goal can be obtained with human control, with machine control, or with mixed control. Only experience, and statistics, will tell.

5.3. Voyeurism As an Antidote to Conflict Escalation

Let us now take a look at the other issues Singer raises relative to the undesirable effects of robotic war: the trivialization of combat, the probable increase of conflicts, a sick voyeurism of our information society, weakening democracies, a growing gap between civilian society and the military apparatus. These issues are all connected and they are not illusory. Were we certain that political leaders of the future would use robotic arms to halt situations of gross injustice, violence,

human rights violations, we would have nothing to fear. When armed militia or the regular army oppress unarmed civilians, children, minorities, then it is likely that intervening political leaders would have the support of public opinion. However, history tells us that political leaders have started wars for much less noble reasons, such as distracting public opinion from internal political problems, or to favour the conquest of new markets on behalf of the economic and financial lobbies that support them – with fake casus belli constructed using mass media controlled by those same lobbies. If one considers that the lack of morality (understood as acting in one's own interest with no respect for others' life and freedom) can nestle also inside the political and economical classes, the alarm – called by the military and civilians interviewed by Singer – seems more understandable. I would worry more about these aspects of the decision process than about the 'weak morality' inherent in a costless military intervention.

As regards the 'porn war,' I think that there is nothing new under the sun. The medium changes (and this is not without its importance), but surely one cannot blame this phenomenon on computers and robots. Think only about Roman gladiators, about the propaganda spread by belligerent nations during the two world wars, and during the cold war, to portray the enemies as inhuman beings who deserve no mercy and one's own weapons as invincible. Of course there are some new psychological elements, but once again we should take a look at human nature rather than trying to solve the problem by banning Predators or footage online.

The porn war that is all the rage on YouTube satisfies a desire for revenge that is inherently human and atavistic. As for the war in the Middle East, it has been fuelled also by insurgents slitting the throat of American prisoners; these have then been picked up and spread online. In other words, the new media have not at all created these instincts from scratch, but they make them visible. It should also be stressed that, while one part of users seem insensitive or even thrilled by looking at such scenes of violence, there have also been reactions of indignation. Therefore these clips – precisely because of their cruel and violent nature – could have also a positive function, because they show public opinion what war really is. By sensitising public opinion, 'war porn' could induce it to take a greater interest in government decisions, and act as a counterweight to the tendency of military interventions to escalate.

5.4. Correct Information As a Counterweight to Alarmism

The new robot prototypes under study, especially those who 'feed' on biomass – the EATR model – have also unleashed ethical discussions. On Fox the news were given an alarmist title: "The Pentagon is working on war robots that feed on dead bodies" [24]. This is false. With famous concern, Robot Technology Inc. and Cyclone – the two companies involved in the project – immediately denied this statement, and clarified that theirs is a vegetarian robot. But despite the clarification the press insisted. In fact, Italian press agency *AdnKronos* reissued both theses, and this with a hyperbolic title: "Here comes EATR, the war robot that can feed on flesh: the debate on cyberethics heats up." The agency's first bulletin is telegraphic:

> Miami – (IGN) – On the battlefield fallen fighters' dead bodies could be the easiest fuel to use for this new robotic weapon that, in order to work, uses biomass. The manufacturing companies deny this: it is 'vegetarian.' But the discussion on the limits to set on these machines, that scientists foresee will soon be able to make autonomous choices, in order to avoid ethical conflicts has been going on for some time [49]

Later a more detailed revision, but alarming nonetheless, was published on the agency's website: "Miami, Aug. 21st, (IGN) – The robots to which we are used today are at best copying dogs or act as vacuum cleaners that run about the house looking for tiny scraps. But, ever faster, 'mechanical creatures' tackle complex tasks, perhaps on the battle stage, as is happening in Iraq and Afghanistan. The latest robot soldier to arrive on the scene, a transport vehicle that moves fuelled by a biomass engine, that is, it burns organic stuff to run, generates some hesitation in the

cybernetic world. Indeed, on the battlefield the most common fuel available might well be human flesh" [49].

The problem of robots eating human flesh is on the desk, even though it is now spoken of as a merely academic hypothesis. After all, it is true that there are dead bodies on the battlefield and it is true that the robots feed on biomass, and since dead bodies are biomass, if one plus one equals two… the rest is consequence.

But perhaps this idea arose in someone's head because of its name? "It is called EATR – which in English sounds uncannily like 'eater'." And yet the makers cannot be clearer. Harry Shoell, manager at Cyclone, puts it thus: "We completely understand the public's concern about futuristic robots feeding on the human population, but that is not our mission," and he adds that no one would dream of violating article 15 of the "Geneva convention" that prohibits the desecration of the corpses of the fallen. The reporter has to take note of this: "The engine developed to power the EATR runs on fuel no scarier than twigs, grass clippings and wood chips" [49].

Yet, can one easily disregard so gluttonous a piece of news? The humanoid cannibal makes splashier headlines than a lawnmower that recycles greens, so it is better to stress the academic hypothesis:

> What would happen, critics ask, if it malfunctioned or ran out of fuel? It would make do with whatever it found, is the answer, and conceive of worrying scenarios along the lines of 'Terminator' or 'Matrix,' science fiction movies where machines take over the planet and use humans as a source of energy [49].

Even though the news is really farfetched, the reporter is right to raise the ethical problem: "In cybernetics the problem of what ethical boundaries should be imposed on these mechanical creations does not go away, given that scientists foresee that very soon it will be possible to make robots able to make largely autonomous decisions." And he cannot avoid quoting Asimov's Three Laws:

> The science fiction writer Isaac Asimov, the author of *I, robot*, had for this purpose conceived three simple laws which, in a remote future, would be programmed into the electronic brains of the automatons. The first of the three, fundamental this one, states: 'A robot may not injure a human being or, through inaction, allow a human being to come to harm.' But he certainly had not taken into consideration the problem of a robot which, in order to exist, might be forced to eat human flesh [49].

Once again, we are not so worried about the actual performance of the machine or the use that one will make of it (it has not yet been used), but the fact that it violates a certain idea of how the world works, to put it like Brown. Ordinary people are convinced that there is a neat, ontological, separation between the animal reign and the vegetal reign, the organic and the inorganic, the living and the dead, the conscious and the unconscious. Robots and GMOs demonstrate that these distinctions are just a convenient heuristic model to classify objects, while reality is much more complex and plastic. A robot can draw energy from his environment and feed himself no more no less like a human being or an animal. With the addition that if there be no potatoes or carrots it can also run on gas or petrol. This worries people because it appears to cast into doubt the uniqueness of humans.

Moreover, the mere existence of EATR conveys that it is at least *technically possible* to build a robot that kills humans and feeds on their flesh, so that it could run for an undetermined length of time. To stop it one would have to switch it off (put it to sleep) or destroy it. If there is no such model it is only because *Robotic Technology* decided to make it vegetarian. Human creative power fascinates some people and frightens others. Hence the ethical controversy. From a pragmatic and rational point of view, it is advisable to serenely accept the 'fact' that the boundaries between organic and inorganic are transient, and strive for these machines to generate more

happiness than unhappiness in the world. Taking it for granted that they don't have feelings (happiness or despair), it would be suitable to give priority to humans and therefore give them the authority to stop the machines at any time in case of malfunctioning or unforeseen and negative collateral effects. However, it seems rational to try also to take advantage of it for civilian or military use. After all, EATR is the ideal lawnmower both as to performance and to save energy. And it would be the only robot able to hinder the action of enemy soldiers or militia over many days in a hostile environment, far from the bases and cut off the system of logistic assistance.

6. Scenario Analysis: Dreams and Nightmares

What will happen tomorrow? If humans rationally tended to choose what is 'good' and to reject what is 'bad,' for themselves and for others, in theory we ought to see a constant improvement of the human condition. But this can only happen in utopias. The problem is that human choices are not always free. They are not always rational. What is good to one group is not always good for another. What is rational at the micro level (individual, social group) is not always rational at the macro level (society, humanity), and vice versa. And finally there is always the possibility of the 'unanticipated consequences of purposive actions,' already studied in detail by sociologist Robert K. Merton [17]. That is, even if we assume social actors to be rational and have positive intentions, there can always be undesired collateral effects. As a popular saying goes: "The road to hell is paved with good intentions."

For the time being, the development of robotics appears unstoppable. We keep hearing that the 21st century will be the century of robots. This happens because, on the whole, such a development appears 'good,' despite the above-mentioned worries and concerns. It appears 'good' also because the classical idea of virtue as a capacity (courage, knowledge, rationality, self-discipline, ability) is once more in favour, and there is no doubt that robots are 'good' in this specific sense. And their 'parents' are every bit as good, since they have been able to transmit to the robots the capacity do to many things. Among these, the ability to fight.

The reason why military applications are being continuously developed is precisely this one: they are 'good soldiers.' First of all they save lives. At the same time they do not have the typically human phobias and weaknesses. In the words of Gordon Johnson of the Pentagon's Joint Forces Command: "They don't get hungry. They are not afraid. They don't care if the guy next to them has just been shot. Will they do a better job than humans? Yes" [43]. Add to this that robots, unlike humans, can be trained and can transmit abilities from the one to the other in an extremely short time: download time. This too is a crucial feature, not just for war, but also in the ever more stringent economical conditions.

At the time of the invasion of Iraq, in 2003, only a handful of drones were in use by the V Corps, the primary command force of the US army. "Today – Singer writes five years later – there are more than 5,300 drones in the US military's total inventory and not a mission happens without them." Therefore, moving on to predictions, one lieutenant of the US Air Force states that "given the growth trends, it is not unreasonable to postulate future conflicts involving tens of thousands" [43].

Between 2002 and 2008, the US defence budget grew 74% to reach 515 billion dollars, not counting some hundred billion dollars, spent on the interventions in Afghanistan and Iraq. Within this expense, the investment into making land unmanned systems is to double every year as of 2001. The Pentagon's order to the constructors is unambiguous: "Make them as fast as possible."

Singer again compares the current situation with that of the industrial take-off, shortly before World War I. In 1908 239 T-Ford cars were sold. Ten years later over a million had been sold. We add that similar situations have been observed with the radio, televisions, computers and telephones. When the home robot boom will take place it will be no less sudden than the technological booms preceding it. The presence of these intelligent machines in homes and in the street will astonish at first, and then be taken for granted.

As regards war machines, one has reached a limit in the development of some manned systems, in particular as regards supersonic aircraft. For example, the intercepting fighter F 16 is too good, in the sense that it is a lot better than the human pilots flying it. It can operate at high speed and follow trajectories, which to a human pilot would be beyond the physically and sensorially endurable. Only a properly programmed computer could maximally exploit the mechanical and aerodynamic features of the latest generation supersonic fighters.

This also goes for other weapons systems. If land robots were able to respond to gunfire, by means of laser sensors and pointers to identify the target, we would see extremely quick responses. Assuming that, in the future, armies on the field will also include robotic soldiers, with the gradual shortening of the loop, then it becomes clear that presence of humans will no longer be possible on the battle field: our reaction times are far too slow.

Therefore humans must inevitably be replaced by robots if the possibilities offered by engineering are to be fully exploited. Bluntly, in the words of one DARPA official, we will have to take into account that "the human being is about to become the weak link in the defence system."

This is why the US are getting ready to set up a "Future Combat System" (FCS), at a total cost of 230 billion dollars, that Robert Finkelstein describes as "the largest weapons procurement in history…at least in this part of the galaxy" [33, p. 114]. The basic idea is to replace tens of thousands of war vehicles with new integrated systems, manned and unmanned, and to write a 34 million lines long software program for a network of computers that will connect all the war machines on land and in the air. Each individual brigade will have more land robots on the field than traditional vehicles, with a ratio of 330 to 300, and one hundred drones under the direct control of ground vehicles. The new robotized brigades could be ready for action in the near future.

Future drones will not necessarily resemble Predator or Reaper. We have already hinted at the futuristic shape of engineering's latest gem, the Northrop Grumman X-47, more resembling a fighter in Battlestar Galactica than a traditional airplane. But also giant drones are under construction. They have a wing span the size of football fields, running on solar panels or hydrogen, capable of being in the air for weeks on end, akin to orbiting spies, but easier to operate. Another direction where research is heading is that of miniaturization, or if we want to use a word more in vogue, that of nanotechnology. In 2006 DARPA gave the green light to a research project with the aim to build a drone with the dimensions and performances of an insect, that is, weighing less than 10 grams, being shorter than 7,5 centimetres, capable of flying at 10 meters/second, with a range of action of one kilometre, and able to hover in the air for at least one minute.

A drone of this kind, other than its military uses, could also be used by the secret services and by the police, to spy or kill. Indeed it could function like smart bombs on a smaller scale. The microdrone would revolutionize all the systems of protection and would have no small consequences on politics and on society. Keep in mind that, in the near future, just as it could be in the hands of the police or the army, the mafia and terrorist groups could have it too. If today it is rather hard and risky for terrorists and mafias to try to kill a politician or some other eminent personality, with the aid of these microscopic flying robots it could become all too easy. It would be enough to remote control them with a SAT-NAV system to reach the victim. The microdrone could thus blow up near the head or other vital organs, or even, alternatively, kill the victim with a lethal injection or with a high voltage electric charge, and then fly off. If an almost invisible nanodrone were to be made, it could enter the nostrils or ears of the victim, killing it with a micro-explosion inside the skull, eluding and confusing the traditional system of protection. Indeed it would not be easy to identify the source, unless one had even more sophisticated electronic systems to monitor and intercept. Setting out to build ever more sophisticated systems of protection, that is, antidotes to nanotechnological weapons, seems therefore more important than putting the weapon itself on the market.

In a hitherto unprecedented situation of vulnerability, it could become all but unsuitable to have a public role in politics, media or entertainment – particularly if such a role is hostile to major powers, mafia or groups with a strong ideological identity. But keep in mind that any 'enlightened lunatic' – laying his hands on this kind of weapons system – could try to kill a famous or powerful

person out of sheer envy or paranoia. Probably, other than systems ID, it would be fitting to prepare a rather rigorous system of traceability that will include satellite systems and systems of land spies able to intercept almost any nanodrone or microdrone in the air or on the ground.

Excessive alarmism could be premature or unfounded, because in history every weapon has had a shield able to stop it. When we went online for the first time and our computers were aggressed by the first viruses, some said that the Internet would never take off as a tool for the masses, because the very expensive hardware could be systematically destroyed by virulent software costing next to nothing. One had not taken antiviruses into account and one had not taken into account the fact that some software would have cost more than the hardware themselves. Of course, more than one user had his computer destroyed by a virus. But these annoying incidents have not taken down the system.

This is to say that the predictions that we are venturing here can only be pure speculation. The future that nanotechnology will generate cannot be foreseen in full. In 2007, when David Leigh, a researcher at the University of Edinburgh, managed to construct a 'nanomachine' the individual parts of which were of the dimension of a molecule, we understood that technology had suddenly projected us into a novel direction with unpredictable consequences. If historical eras are defined by materials (stone, copper, bronze, iron, plastic, etc.), then we have entered into the age of nanomaterials [28]. What will it bring us? Leigh could not tell: "It is a bit like when stone-age man made his wheel, asking him to predict the motorway" [33]. We have entered into a new world, but it is simply impossible to know which kind of world it will be. Any presumption to do so will therefore miss the mark.

The future will be a world of nanomachines, but also the world of androids. An android (or a humanoid) is a robot resembling a human and able to imitate many human behaviours; many designers hope that they will also be able to think and feel in ways analogous – even though not absolutely identical – to those of humans. Ian Pearson had defined 'androids' as machines that have a consciousness, linking the concept not so much to the anthropoid shape, as to the anthropoid mind. Scientists and engineers are already designing humanoid soldiers [18].

The military hopes that androids – whatever is meant by them – will be even better warriors than humans. When DARPA asked the military and scientists to indicate what role robots will play alongside humans, and then without them, in the near future, they replied in the following order: demining, recognisance, vanguard, logistic, and infantry. Oddly, air defence and driving vehicles, where their use is common, were mentioned only at the end. When they were asked to give a date when it will be possible to send humanoid robots to the battlefield instead of infantrymen, the military said 2025 and the scientists 2020. Robert Finkelstein, president of Robotic Technology Inc., finds these forecasts too optimistic and gives 2035 as the date when androids will first be sent to the front. In any case it is not a long time. Many readers of this book will still be among us to verify the prediction.

7. Conclusions

Since the world began, wars have been fought by 'mixed' armies under various flags: an alliance of humans, animals and machines on the one hand, against an alliance of humans, animals and machines on the other. This was the case in the days of Alexander the Great and it is the case today. The war machines that Archimedes or other Hellenistic engineers conceived are not as powerful as the robots we today send out to the battlefield, but still they are their cultural ancestors [5, pp. 124-130]. To wonder if the Golem model will arrive is like asking: will this pattern change? The on-going emergence of sophisticated objects that violate ordinary people's expectations as to how the world works or should work leads one to suspect that war as a whole could also yield some surprises. The greatest fear is that of seeing, for the first time in history, homogenous and no longer mixed deployments, namely: machines against humans. Science fiction and apocalyptic journalism insist on this matter.

Any prediction, even when founded on rigorous studies of trends, always have a large speculative component by virtue of the complexity of the system. All the same, scenario analyses are still useful and therefore we will not shy away from venturing a prediction. All our analyses lead to the conclusion that the hypothesis of a 'species' war between humans and machines, ending with the defeat of the former, is highly unlikely in the 21st century. The reasons underlying this belief are all in all six.

1) Metaphysical Uncertainty. One must consider first of all that it might be impossible for human consciousness to understand itself or replicate by scientific means. Even though materialistic metaphysics has shown itself most fecund to science in the last few centuries, and thereby made a privileged hypothesis, this does not allow us to exclude with absolute certainty the plausibility of idealistic or dualistic metaphysics. If the supporters of dualistic mind-matter ontology – like Pythagoras, Plato, René Descartes, Karl Popper, etc. – are correct, then robots can never be conscious in the same way as a human being.

2) The Complexity of Consciousness. Even if we postulate that materialistic metaphysics is correct, it is necessary to acknowledge how *hard* our task is. There has been remarkable progress in Logic, Computer Science, Psychiatry, Biology and Philosophy of Mind in the last centuries, but we are still a long way from understanding the concept of consciousness. And we cannot replicate what we do not understand. We can only make something different. In addition, considering that we have not yet managed to solve technical problems that are apparently simpler, such as a cure for baldness or caries, it is understandable that some previsions about the technological development of androids are regarded as overly optimistic.

3) The Alien Character of Artificial Consciousness. Even if we postulate that consciousness is just an emerging property of matter when suitably organized, and admit that artificial consciousness could emerge as an undesirable collateral effect from other actions, this does not imply that alien intelligence would necessarily be a hostile artificial intelligence. In other words, even if our machines were to spontaneously acquire their autonomy for reasons beyond our comprehension, this does not logically entail that they will be violent towards us. We tend to view humans as angels and machines as potential Terminators, but all anthropological and biological observations demonstrate that it is man in fact who is the most dangerous and aggressive predator produced by evolution. An alien intelligence could be benevolent precisely because it is alien, and not in spite of it. In other words, the alien character of artificial intelligence is in reality an argument against it being hostile. This is how things stand now until proven otherwise.

4) Potency of Technological Man. Even if a hostile artificial intelligence were to emerge, even if our robots were to rebel against us, humans are still powerful enough to engage in the equivalent of an 'ethnic cleansing' of the machines. Let us not forget that humans would not be fighting the robots with bows and arrows, but with blinded tanks, airplanes, remote controlled missiles, and, in extreme cases, nuclear devices. The battle would be between two hitherto unseen armies: on the one hand an alliance of manned systems and unmanned systems that have remained faithful to humans, and on the other hand unmanned systems remote controlled by hostile artificial intelligence. The final outcome of this hypothetical clash is anything but certain.

5) Evolution of Technological Man. Even if unmanned systems were to evolve to the point of becoming more potent than any manned system, we should not forget that humans themselves will presumably undergo an evolution by technological means. Humans, using genetic engineering or hybridising with machines via the implants of microchips in the brain or under the skin, could cease to be the weak link in the chain. In the future they might react at a thinking level equal in speed and precision to those of machines.

6) Man-Machine Hybridization. Finally we must consider that, because of technological development in the fields of bioengineering and of robotic engineering, we might never have a conflict between the organic and the inorganic worlds, between humans and machines, between carbon and silicon, simply because there will be a real and true ontological 'remixing.' There will be human beings empowered with electro-mechanical parts and robots with organic portions in their brain. Therefore it is not ontology that will decide the alliances.

In conclusion, we believe that in the 21st century there will still be humans, machines and animals serving under one national flag, waging war against humans, machines and animals serving under another national flag. When this system has disappeared, if there are still conflicts, in our opinion it will be more likely to see a variety of sentient beings (humans, transhumans, and posthumans) on the one hand, under one flag, against a variety of sentient beings (humans, transhumans, and posthumans), under another flag. But we are speaking of a very remote future.

The more concrete and pragmatic recommendation that I would now give makers of robotic weapons and their political and military customers is to always work on parallel projects, conceiving, for each robotic weapon that they construct, another weapon able to control and destroy it. This precaution could reveal itself useful both in the science fiction scenario of the emergence of hostile artificial intelligence, and in the more prosaic and plausible scenario that the robotic weapon falls into enemy hands.

However I believe it inopportune and irrational to apply the maximalist version of the precautionary principle. By maximalist version I mean an interpretation of 'precaution' that would mean banning any technology that does not present itself as absolutely risk-free.[4] First of all, there is no technology or human action that is risk-free, because it is not possible to foresee the whole range of future developments inherent in a certain choice. As it is said, the flapping of a butterfly wing in the Southern Hemisphere can cause a hurricane in the Northern Hemisphere. Second, since we do not live in paradise and since processes pregnant with a future that we do not know are already in the making, non action does in no way guarantee that the results will be better for our group or for all humanity. This to say that the failure of the butterfly wing to flap in the Southern Hemisphere could also provoke an extremely serious drought in the Northern Hemisphere. Finally, the precautionary principle (at least in its maximalist interpretation) never pays sufficient attention to the benefits that might derive from a risky action. On closer inspection, fire has been a risky undertaking for Homo Erectus. During the million or maybe more years that separate us from the discovery of the technique of lighting and controlling fire, many forests and cities have been consumed by flames because of clumsy errors by our ancestors. Billions of living beings have probably died because of this technology. And today we still hear of buildings that burn or explode, causing deaths, because of malfunctioning central heating systems or mere forgetfulness.

Yet what would humans be without fire? If our ancestors had applied the maximalist precautionary principle, rejecting fire because it is not risk-free, today we would not be Homo Sapiens. This dangerous technology has indeed allowed us to cook our food and hence for hominid jawbone to shrink, with the ensuing development of language, and of the more advanced idea of morality and technology that language allows. In brief, today we would not even argue in favour or against the precautionary principle, or indeed any principle, because these require language for their formulation.

References

1. Ahmed, A. *The Thistle and the Drone. How America's War on Terror Became a Global War on Tribal Islam*, Washington D. C.: Brookings Institution Press, 2013.
2. Bashir, S., and R. D. Crews (eds.). *Under the Drones. Modern Lives in the Afghanistan-Pakistan Borderlands*, Cambridge (MA) and London: Harvard University Press, 2012.
3. Brown, A. War crimes and killer robots, *The Guardian*, March 18, 2009.
4. Campa, R. Kodeksy etyczne robotów: zagadnienie kontroli sprawowanej przez człowieka, *Pomiary Automatyka Robotyka* 3/2011, pp. 86-90.
5. Campa, R. *La rivincita del paganesimo. Una teoria della modernità*, Monza: Deleyva Editore, 2013.
6. Cappella, F. Big Dog, *Neapolis*, March 20, 2008.
7. Chomsky, N., and A. Vltchek. *On Western Terrorism. From Hiroshima to Drone Warfare*, London: Pluto Press, 2013.

8. Clapper, J. R., J. J. Young, J. E. Cartwright, and J. G. Grimes. *Unmanned Systems Roadmap 2007-2032*, Department of Defense (USA), 2007.
9. Clapper, J. R., J. J. Young, J. E. Cartwright, and J. G. Grimes. *Unmanned Systems Roadmap 2009-2034*, Department of Defense (USA), 2009.
10. Devereux, T. *La guerra elettronica: arma vincente 1812-1992*, Varese: Sugarco Edizioni, 1993.
11. Evangelista, M., and H. Shue (eds.). *The American Way of Bombing. Changing Ethical and Legal Norms, from Flying Fortresses to Drones*, Ithaca and London: Cornell University Press, 2014.
12. Feletig, P. Robot per mare, per cielo e per terra ormai in guerra si va senza uomini. *Repubblica*, February 1, 2010.
13. Ishihara, K., and T. Fukushi. Introduction: Roboethics as an Emerging Field of Ethics of Technology, *Accountability in Research* 17 (6): *Roboethics*, 2010, pp. 273-277.
14. Kopacek, P. Roboethics, *IFAC Proceeding Volumes* 45 (10), 2012, pp. 67-72.
15. Maruccia, A. Era del Robot, umanità in pericolo, *Punto Informatico*, February 28, 2010.
16. Meggiato, R. EATR il robot vegetariano. Sviluppato il primo robot in grado di alimentarsi da sé con dei vegetali, *Wired*, July 16, 2010.
17. Merton, R. K. The unanticipated consequences of purposive social action, *American Sociological Review* 1, 1936, pp. 894-904.
18. Nath, V., and S. E. Levinson. *Autonomous Military Robotics*, Heidelberg: Springer, 2014.
19. Nichols, M. Italian firm to provide surveillance drone for U.N. in Congo, *Reuters*, August 1, 2013.
20. Northrop Grumman. X-47B UCAS Makes Aviation History…Again! Successfully Completes First Ever Autonomous Aerial Refueling Demonstration, retrieved from: www.northropgrumman.com (May 13, 2015).
21. Page, L. US war robots in Iraq 'turned guns' on fleshy comrades, *The Register*, April 11, 2008.
22. Petroni, A. M. Liberalismo e progresso biomedico: una visione positive, In R. Campa (ed.), *Divenire. Rassegna di studi interdisciplinari sulla tecnica e il postumano, Vol. 2*, Bergamo: Sestante Edizioni, 2009, pp. 9-43.
23. Poundstone, W. *Prisoner's Dilemma*, New York: Doubleday, 1992.
24. R. Z. Tecnologie inquietanti, il Pentagono studia robot da guerra che si nutrono di cadaveri, *Tiscali Notizie*, July 17, 2009.
25. Relke, D. M. A. *Drones, Clones, and Alpha Babes. Retrofitting Star Trek's Humanism, Post-9/11*, Calgary: University of Calgary Press, 2006.
26. Rogers, A., and J. Hill. *Unmanned. Drone Warfare and Global Security*, London: Pluto Press, 2014.
27. Sanchez, M. Robots Take Center Stage in U.S. War in Afghanistan, *Fox News*, March 23, 2009.
28. Serreli, V., C. Lee, E. R. Kay, and D. Leigh. A molecular information ratchet, *Nature* 445, February 1, 2007, pp. 523-527.
29. Sharkey, N. Robot wars are a reality. Armies want to give the power of life and death to machines without reason or conscience, *The Guardian*, August 18, 2007.
30. Sharkey, N. Ground for discrimination. Autonomous Robot Weapons, *RUSI Defence Systems*, October, 2008, pp. 86-89.
31. Sharkey, N. The Ethical Frontiers of Robotics, *Science* 322 (5909), December 19, 2008, pp. 1800-1801.
32. Singer, P. W. Robots at War: The New Battlefield, *The Wilson Quarterly*, Winter 2009.
33. Singer, P. W. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, New York: The Penguin Press, 2009.
34. Sloggett, D. *Drone Warfare. The Development of Unmanned Aerial Conflict*, Barnsley: Pen & Sword Aviation, 2014.
35. Sparrow, R. Killer Robots, *Journal of Applied Philosophy* 24 (1), February 2007, pp. 62-77.

36. Sullins, J. P. An Ethical Analysis of the Case for Robotic Weapons Arms Control, In Podins, K., Stinissen, J., and Maybaum M. (eds.), *5th International Conference on Cyber Conflict*, Tallinn: NATO CCD COE Publications, 2013.
37. Turse, N. *The Changing Face of Empire. Special Ops, Drones, Spies, Proxy Fighters, Secret Bases, and Cyber Warfare*, New York: Dispatch Books, 2012.
38. Tzafestas, S. *Roboethics: A Navigating Overview*, Springer, Cham 2016.
39. University of Sheffield. Killer Military Robots Pose Latest Threat To Humanity, Robotics Expert Warns, *ScienceDaily*, February 28, 2008.
40. Veruggio, G. The EURON Roboethics Roadmap, "Humanoids'06": IEEE-RAS International Conference on Humanoid Robots, December 6, 2006, pp. 612-617.
41. Veruggio, G., and F. Operto. Roboethics: Social and Ethical Implications of Robotics, In B. Siciliano, and O. Khatib (eds.), *Springer Handbook of Robotics*, Berlin Heidelberg: Springer, 2008, pp. 1499-1524.
42. Weinberger, S. Armed Robots Still in Iraq, But Grounded, *Wired*, April 15, 2008.
43. Weiner, T. New Model Army Soldier Rolls Closer to Battle, *New York Times*, February 16, 2005.
44. Whittle, R. *Predator. The Secret Origins of the Drone Revolution*, New York: Henry Holt and Company, 2014.
45. Winnefeld, J. A., and F. Kendall. *The Unmanned Systems Integrated Roadmap FY2011-2036*, Department of Defence (USA), 2011.
46. Winnefeld, J. A. and F. Kendall. *The Unmanned Systems Integrated Roadmap FY2013-2038*, Department of Defence (USA), 2013.
47. Wittes, B. and G. Blum. *The Future of Violence. Robots and germs, Hackers and Drones*, New York: Basic Books, 2015.
48. Zaloga, S. J. *Unmanned Aerial Vehicles. Robotic Air Warfare 2017-2007*, Oxford and New York: Osprey Publishing, 2008.
49. Creato EATR, robot da guerra che può nutrirsi di carne: s'infiamma il dibattito sulla cyber-etica," *AdnKronos*, 2009, August 21.

Notes

1. The United States is one of the most active nations at the cutting edge in the development and use of these technological products.
2. *Rocket Propelling Grenade* – a Soviet manufactured anti-tank grenade launching system.
3. Originally elaborated by Merrill Flood and Melvin Dresher at the RAND in 1950, the prisoner's dilemma was later formalized and given its present name by Albert W. Tucker [23].
4. On this problem we invite the reader to have a look at Petroni's *Liberalismo e progresso biomedico: una visione positiva* [22]. Even though he mainly focuses on biotechnologies, the article offers a detailed and convincing analysis of the precautionary principle.

Studia Humana
Volume 8:1 (2019), pp. 49—66
DOI: 10.2478/sh-2019-0003

How an Advanced Neurocognitive Human Trait for Religious Capacity Fails to Form

Margaret Boone Rappaport

The Human Sentience Project,
Tucson, Arizona USA

e-mail: msbrappaport@aol.com

Christopher Corbally

The Human Sentience Project,
Tucson, Arizona USA,
Vatican Observatory and University of Arizona
Dept. of Astronomy, Tucson, Arizona USA

e-mail: corbally@as.arizona.edu

Abstract:
The authors present an evolutionary model for the biological emergence of religious capacity as an advanced neurocognitive trait. Using their model for the stages leading to the evolutionary emergence of religious capacity in *Homo sapiens*, they analyze the mechanisms that can fail, leading to unbelief (atheism or agnosticism). The analysis identifies some, but not all types of atheists and agnostics, so they turn their question around and, using the same evolutionary model, ask what keeps religion going. Why does its development *not* fail in one social group after another, worldwide? Their final analysis searches for reasons in important evolutionary changes in the senses of hearing, vision, and general sensitivity on the hominin line, which together interact with both intellectual and emotional brain networks to achieve, often in human groups, variously altered states of consciousness, especially a numinous state enabled in part by a brain organ, the precuneus. An inability to experience the numinous, consider it important, or believe in its supernatural nature, may cleave the human population into those with belief and those with unbelief.
Keywords: agnosticism, atheism, cerebellum, cognitive evolution, *Homo erectus*, *Homo sapiens*, neuroscience, numinous, parietals, precuneus.

1. Uniformity with Variation

Much has been written over the years that tries to define atheism and different forms of it. The best conclusion we can draw from this voluminous literature is that "unbelief," or what we would define as a disbelief in the supernatural, appears in all human societies. Examples come most easily from the so-called Great World Religions – Judaism, Christianity, Taoism, Buddhism, and Islam – because they are well documented.

When we look at non-modern societies, we are mainly convinced of a wide range of religious expression that all other humans appear able to identify as "religious" behavior. Members of some societies are fully involved in religious activities and appear to believe in the supernatural; members of other societies are barely concerned with religious observances, and their range of belief remains unknown [3]. It is precisely in some tribal societies that religious behavior appears, at first, to be most uniform: Everyone participates, everyone goes through rites of passage, everyone seems to believe in the supernatural. And yet, ethnographies routinely capture the contrarian individual who refuses to go along, who leaves and stakes out a new home with a lover he or she should not have, or who simply does not participate and assumes a non-traditional role, like the berdache in a Native American culture who refuses to go along with the entire male role [25]. That non-conforming behavior almost always extends to some change in religious participation, which is tied intimately to conformity. Non-conformity suggests the possibility of unbelief.

When we examine the literature on religion, both modern and pre-modern, we conclude that while some religious behavior appears as a constant from society to society, its level of expression, fervor, or depth is quite varied, from a deep faith in, to an unbelief in the supernatural. Religious expression appears in all known human societies according to the ethnographic and historical literatures, and apparently always has, from the time members of our genus Homo first had religious thinking and engaged in religious behavior. In spite of the uniformity of appearance of some religion in every society, there can be wide variation among individuals. These conclusions are found in surveys of so-called "religiosity" in modern societies [30], [31]. Uniformity with variation in modern *Homo sapiens* is a judicious place to begin looking at the opposite of religious thinking – atheism and agnosticism. We hinge our analysis on "religious thinking" within a theoretical framework of cognitive archaeology [13]. In modern societies "religious behavior" can be faked and in very early prehistoric societies, there are few finds that suggest religion. We approach religious thinking as a neurocognitive trait that emerged in our evolution.

The logic of our approach is to understand how religious capacity emerged in an evolutionary context, to speculate on how it may fail to emerge in full or fractional measure in living human beings, and then to ask whether we have logically defined all types and degrees of unbelief. Much of our logic is driven by the burgeoning science of human genomics, which shows that very few biological traits are determined by single genes. Human traits – both physical, and the more complexly derived cognitive traits – are driven by multiple, interacting genes. That variety of genes and their potential for variable, phenotypic expression drives the possibility that atheism (or degrees of it, as in agnosticism) could be based upon the expression of many genes and behaviorally, at varying levels of intensity, or not at all, i.e., unbelief. Therefore, this stands as a fertile field for research in the future. At the present, only a very few genes affecting religious thinking have been identified.

2. Logic, Hypothesis, and Question

In this paper on atheism, or unbelief, we have chosen to analyze the "failure to form" a human biological trait that is relatively common, but far from phenotypically uniform. This failure can partially explain the degree or level of religious behavior. However, since religious behavior is driven by social factors apart from biology, there is never a clear one-to-one correspondence between biology

and behavior. And, there is individual, intellectual choice in humans about expressing the trait of religious capacity. Someone may have "the full monty" for religious thinking, but not express it behaviorally. Conversely, another person may have the religious trait to a modest degree but be heavily involved in religious activities.

Therefore, we hypothesize that the biological trait of religious capacity ("religious thinking") has natural variation and at the far end of the scale of variation is atheism. Elsewhere, we have proposed that religious capacity is one of the most complex neurocognitive traits that modern humans possess, and that it makes use of a large number of brain capacities enabled by many brain networks, including, for example, the combined action of the fronto-parietal and the default networks [15], and neural connections between the cerebral cortex and the cerebellum described below [41], [43]. Yet, these are but two of the networks and one pathway involved in religious thinking and theological creativity (defined as the creation of new knowledge using a theory of the relationships between humans and the supernatural that is consistent with broader cultural themes). There are many other neurological capacities involved in religious thinking and participation, including the entire range of brain capacities that manipulate cosmological concepts visuospatially. These include cognitive capacities to imagine and manipulate the self and supernatural beings within those spaces, which are seated, in part, especially in the precuneus [6], [7], [8], [9], [10], [11], [22], [40], [42], [48].

Even with knowledge of only a few genes and several networks at this early point in time, our conviction is that complex combinations of genes and resulting brain capacities routinely produce religious thinking. Religious capacity's emergence – either in evolution or individual ontogeny – is not simple. There is plenty of room for variation and for failure. There may well be *increasing* variation for the highest, most advanced, human neurocognitive traits because of the increasingly complex genomics underlying them. When we recall that apparently many of the networks used in higher cognitive processes are "exapted" [12], that is, are re-used and re-worked from their original neurological functions, we can see that the process of producing "religious capacity" in a single individual has much room for variation and failure. This is a difference in ontogeny, certainly not a moral failure.

The evolutionary "Building Blocks" identified below make religion not just an idea, but part of our biology through processes whose origins are ancient and overlapping. There is a long history for the biological trait we call "religious capacity" and it has natural variability of phenotypic expression. Some people have what seems like "deep faith," while others appear to have less conviction (whether they behave according to religious principles or in religious activities, or not). Some people show striking creativity in their theological treatises, and others barely connect theology to their worshipful activities. As noted, a "theology" is a theory about the relationships between humans and the supernatural. Theological creativity is not necessarily co-incident with religious thinking, although they tend to overlap, especially for religious leaders from shamans to prelates, who can be quite creative in their interpretations of daily events, dreams, social conflicts, and individual motivations.

The contention that religious capacity is a highly complex neurocognitive trait with broad phenotypic variation is consistent with our hypothesis that it is an evolved biological trait. Many human biological traits are widely held and variably expressed, but some humans do not have them, at all. There are also traits that can be traced to the genomic level, which only certain proportions of humans have. Not everyone has blue eyes. Not everyone is equally hirsute. Not everyone is a "highly sensitive person" [1]. As we learn more and more about the human genome, we see that biological traits can have many separate genes that affect their phenotypic expression. This is solid biology. If religious thinking is one of the most complex, if not the most complex, human neurocognitive trait that now characterizes our species, then we would not be surprised by its *absence* in a proportion of humans. Biological development of individuals goes awry, with wonderful to tragic results.

We propose that what adult humans[1] perceive as "atheism" is the absence of religious thinking to a noteworthy degree. Some individuals stand out as particularly unbelieving. They may expound on their views widely or be very quiet about it. They may participate in some types of religiously

sponsored events, and yet still have unbelief. When recognized, they can encounter an entire range of reactions from exclusion to congratulations. In most modern industrialized societies, atheists are generally accepted socially for most purposes and not excluded from social activities. They join voluntary associations of others who have unbelief. In tribal, non-modern, and evangelical groups, the views of unbelievers can lead to expulsion and even death.

To define those *without something* therefore hinges upon what capacities others do have, again to various degrees. It also opens the door to the contention that many have made concerning atheism – that it somehow causes those humans who exhibit unbelief to see "more clearly" and "more realistically" than others who are "afflicted" or "burdened" by religious thinking. This follows the dictum that religion is the "opium of the people" [26] and that the future for humans involves a clearer perspective that is somehow "free of religious thinking." This logic is quite contrary to religious doctrine, of course, but it is quite cogent. Its opposite is also quite cogent – that religious capacity allows those who have it to "see things" that others cannot fathom, again quite logical.

Let us look at the evolutionary progression that we propose for religious capacity's emergence, and then use that framework to analyze how it could possibly fail in an individual. An analysis of atheism as a failure to form leads unexpectedly to conclusions about religious capacity's fundamental nature, why it is successful in supporting the social group, and why it has become so bound up with social control.

3. An Evolutionary Model for Religious Capacity Suggests Ways It Can Fail to Form

We summarize a theoretical model whose foundation lies in research details published previously [34–38]. This model traces the biological foundations of religious capacity as a biological trait, and the reader is encouraged to see these papers and the source studies we relied upon. Religious capacity's antecedents are all evolutionary innovations without which today's biologically based capacity would not exist. Some are well known, like Primate sociality, and others are just being discovered, like the importance of the expansion of the cerebellum in higher Primates. Our task was to configure a model based on circumstantial evidence in cognitive science, neuroscience, and genomics, and to project these findings backward to see how they articulate with evidence from paleobiology and "stones and bones" archaeology – always the knowledge base that determines the outlines of our evolution, as we currently understand it. That understanding will surely change, with new discoveries, but we believe this is a good, first attempt. We know of no comparable analysis.

We ask: How did the first and later Primate species need to evolve in order to emerge, 65 million years later, with something like religious capacity in *Homo sapiens*? When findings from traditional archeology, the new cognitive archaeology, primatology, paleoneurology, cognitive science, neuroscience, population genetics and the burgeoning field of human genomics (of extinct and modern forms) point to the same types of changes, the biological foundation for religion seems more certain. True, it probably did not fully emerge until our species, *Homo sapiens*, but there were antecedents. The sequence of evolutionary breakthroughs we identify makes religion not just an idea or a cultural fabrication, but part of our biology. Each step through evolutionary time adds a needed biological basis for modern human religious thinking, which is cognitively, emotionally, and intellectually very complex, but is uniform in its support for the social group. Religious capacity may rely on hundreds of brain capacities, if indeed there are thought to be "thousands" [16].

It should be noted that Neanderthals are compared and contrasted with our species repeatedly. Both paleoneurological [6-9] and very new findings on the cerebellum [41], [43] suggest to us that Neanderthals, in all likelihood, did not have religious thinking like *Homo sapiens*. The archaeological record on Neanderthals remains mixed, and open to conflicting interpretations, although recent genomic studies point to important cognitive deficiencies in Neanderthals, when compared to modern humans [17], [19], [29], [39], [47].

We have recently added a new Building Block to our model, which follows the very first one, so it becomes Building Block 2, below. Because this is our first presentation of Building Block 2, we briefly summarize our rationale for inserting it into the original model. Intelligence, as it is determined by the unusual (but not completely unique) expansion and reorganization of the lateral cerebellum, represents an evolutionary innovation, like others, without which religious thinking would not exist today. It is fundamentally an upgrade in sheer computing power plus an ability to store internal models of external models in the ape and human cerebellum, and, for humans, to connect them to expanded association areas in the cerebral cortex. An expansion of the lateral cerebellum is novel to mammals, and for humans, it is connected to "higher cognitive functions" [41].

Our newest Building Block reflects convergent evolution that occurred in three different Orders of the Class Mammalia, including ours: the primates, cetaceans, and pinnipeds (seals). Smaers and colleagues [41] point out that changes in the lateral cerebellum are more reflective of the modularity and interconnectivity necessary for intelligent behavior, than measures of sheer volume would be. Their analysis suggests that cognitive capacities are "scaffolded" by modifications in the mammalian cerebellum, but they only fully occurred in these three Orders of Mammalia. In our Order Primates, it occurred significantly in the anthropoid apes, which gave rise to humans. The authors delve deeply into a statistical proof that lateral cerebellar expansion is strongly related to other measures of intelligence and complex communication. Tanabe and colleagues confirm this interpretation, noting that the cerebellar "neuroanatomical organization may affect innate learning, cognitive ability, and the human capacity to innovate" [43]. We logically connect the ability to handle complex information, and innovate from it, to religious thinking and theological creativity, in fact, all forms of creativity, including art and science.

The size of cerebellar units, which function something like computer chips, is directly related to the number of internal models that humans can store, and the connections to the association areas of the cerebral cortex. Because of these connections to the cerebral cortex, the cerebellum is therefore connected to the primate fronto-parietal network – a feature that is not observed in other mammals. Therefore, the cerebellum is fundamentally involved in human thinking about the supernatural: cosmological space, supernatural beings, and imagining the self in interaction with these beings, in these spaces. Theological creativity uses models to create new stories to illustrate religious teachings, and new religious tenets emerge to summarize these stories. We propose that all these features of religious thinking articulate with the substantial and more general human capacity to manipulate visuospatial information (real or imaginary), especially in the parietal lobes. Three-dimensional imagination is central to physics and to theology, although both rely on many more human brain capacities, too.

Cerebellar re-organization comes after our first foundational Building Block, which represents a more general feature of all Primates (not just the anthropoid apes) – sociality. We place re-organization of the lateral-medial cerebellum *after* (in evolutionary terms) primate sociality and *before* the emergence of the first true ape, Proconsul (Building Block 3). Various ranges of dates are given for Proconsul, from 23-25 mya [million years ago] to 14-23 mya. The more general term, proconsulids (representing 10 different genera) date 17-22 mya. Our sequence of Building Blocks assumes that cerebellar-cortical reorganization was ongoing somewhat before Proconsul fully evolved, and that the species stabilized around our estimated date for Proconsul of 19 mya.

The evolutionary emergence of religious capacity in the genus Homo relied upon the following 10 Building Blocks:

Building Block 1. Sociality in all primates, 65 - 55 million years ago.

Building Block 2. Reorganization of the lateral-medial cerebellum in the anthropoid apes, Order Primates, leading to modularity and increased capacity to store internal models, and, interconnectivity between the cerebellum and the association areas of the neocortex [41]. The result was "intelligent

behavior" and ability to innovate [41] [43]. We see this ability to innovate in science, art, and religion of *Homo sapiens*.

Building Block 3. A basic ape model from the Miocene, beginning around 19 million years ago, with Proconsul, the first true ape.

Building Block 4. Realignment of the senses, with upgrades of vision and hearing on the line to humans and some modern apes.

In some groups of the ancestral ape population giving rise to the genera Homo and Pan in Africa, Building Blocks 5 - 9 emerge:

Building Block 5. Lengthening developmental trajectory or "secondary altriciality" and the downregulation of aggression, 8 - 10 million years ago.

Building Block 6. Greater social tolerance among adults, especially while feeding.

Building Block 7. Further upgrades in intellect to help to manage aggression in the social group.

Building Block 8. Greater sensitivity emerges, both general sensitivity (in terms of heightened awareness and preparation for action), and sensitivity that engages the emotions.

Building Block 9. Biological foundations for culture emerge in ancestors to both Homo (strongly) and Pan (weakly). The first evidence for culture in our genus was in *Homo habilis*, who made the first stone tools found in the archaeological record. Culture also likely characterized earlier, bipedal Australopithecines, who did not have shaped-stone tool traditions, but probably used stones to butcher.

Moral and religious capacities emerge relatively late – moral capacity in *Homo erectus* and religious capacity in *Homo sapiens*.

Building Block 10. Moral capacity emerges in *Homo erectus*, 1 - 1.5 mya, after the species controls fire and a learning context called "The "Human Hearth" develops. The reader is referred to the full theoretical development of this model [32-33], which includes cognitive features that characterize rudimentary morality and gives research findings that support their presence. Phenotypic expression of moral capacity in *Homo erectus* (and later in *Homo sapiens*) include all of the following at the same time:

a. A mental step both back and up
b. An arbitration mechanism that operates along a timeline
c. An evaluation using a valence from good to bad
d. A regretfully dispassionate reasoning
e. A tentativeness in a mental balancing act
f. A sad rejection of "wantonness"
g. A capacity for empathy with someone receiving moral judgment
h. The experience of a burden
i. Resolution on the part of the group
j. Hope and faith in the future on the part of the group

Our model includes the emergence of religious capacity in *Homo sapiens*, stabilizing at around 120,000-130,000 years ago, according to studies on globular brain shape of fossil skulls [34-36]. Skulls began to round in a manner typical of the more modern human species before this, by around 300,000 year ago, according to finds from Jebel Irhoud, Morocco [23]. The human skull began to round in response to the expansion of brain tissues beneath, particularly the precuneus, part of the parietals [6-9], but also due to the enlargement of the cerebellum and other underlying brain tissues [19]. In this latter study, there is evidence that the cerebellum was an important difference between *Homo sapiens* and *Homo neanderthalensis*.

We understand "theological creativity" as part of the development of religious capacity in *Homo sapiens* alone. Through time, theologies were shaped increasingly by various cultures so that they are consistent with other cultural themes. However, they maintain a remarkable number of fundamental similarities cross-culturally, and despite their differences, testify to a single neurocognitive origin for religious capacity and theological creativity in our species. Our model

includes the earlier emergence of moral capacity in *Homo erectus*, but not religious capacity. Our view is that *Homo erectus* may have had chanting, percussion, and storytelling, but only in *Homo sapiens* is there music, a capacity enabled, again, by the precuneus in the modern human parietal lobes. Furthermore, modern humans are the only ones to have internally consistent and structured theories about the relationships between humans and the supernatural. Religious and moral capacities are usually intimately joined in modern humans, so what is true for moral capacity remains true for religious capacity. Yet, even now, moral and religious capacities are separable, both theoretically and practically, as many types of organizations involve moral thinking but are not specifically religious.

The fact that moral and religious capacities can be conceptually teased apart is particularly important for our task related to atheism, i.e., identifying mechanisms whereby religious capacity fails to form. If religious and moral capacities are separable, moral capacity can be present without necessarily its frequent conveyor (religious capacity), and that makes sense from what we know of the modern atheist: They are "without God," so to speak, but not necessarily without morality, as with "ethical humanism." Before we go on further to address the nature of religious capacity and its absence, let us first look at which Building Blocks suggest mechanisms that might fail, and therefore produce unbelief or atheism.

4. Mechanisms Whereby Religious Capacity Can Fail to Form

We now use our evolutionary model as an analytical framework to discover possible ways in which religious capacity could fail to form.

Building Block 1: Primate Sociality

There are myriad ways in which social development can affect a proclivity for unbelief or atheism. Parents and family members may profess little faith or choose not to participate in religious activities. The question then becomes whether an assumption of unbelief by a family member results from inheritance, learning, or both (the nature/nurture question). More fundamental for the adult atheist might be either an event that encouraged or signaled unbelief, or a conscious choice – an intellectual conclusion – that unbelief made the most sense. Then, there would be a choice as to whether to accept the social consequences of unbelief, especially if they are onerous. Unbelief would become a life choice, like political party, and it would be shorn up by others characterized by unbelief.

On one end of the sociality scale are individuals who are developmentally delayed and not intellectually capable of participating in the society's usual religious activities. This can result, for example, from a social anxiety disorder or a genetic disorder along autism-schizophrenia scales that prevents the more usual, mutually satisfying, and effective social communication. This also includes some very intelligent individuals with severe autism who are disabled in terms of social communication. Developmentally delayed and autistic humans are usually not be able to comprehend basic tenets of a religious creed, its logic, ethics, lifeway, or its supernatural beings. While some creeds identify the disabled as "touched by the supernatural," this appears to be a culturally fabricated explanation, perhaps to soften a sense of helplessness or fear. The analysis in this paper strongly suggests that humans with very low intellect are rarely able to participate either socially or intellectually in religious thinking or religious life. Indeed, their inability provides support for the heightened intellect needed for religious thinking that is found on the evolutionary line to modern *Homo sapiens*.

For our evolutionary model, we are reminded that religious capacity tends to involve a person with other people, in social groups and in whatever rituals are required and whatever social events mark the calendar. Religious capacity does not quite make sense except in a social context, as part of a social institution and support for it, and, as the result of social learning. True, there is the lone monk

worshipping by himself, but his doctrine and usually his rituals rely on a belief system fashioned by others. We conclude that primate sociality – so old and so fundamental to how we live our lives – is *sine qua non* for religious capacity. So, what of the humans who have no belief? They can be left stranded socially, or they can be free to choose another social group in modern societies. That freedom is appealing to some people who evidence unbelief. Atheism does have the effect, especially where there are other choices readily available, of "freeing" an individual from some social strictures, but not all.

Building Block 2: Primate Reorganization of the Lateral-medial Cerebellum

There are many different ways that intelligence can affect unbelief or atheism. Many of those who write on atheism make cogent arguments for atheism, and it is clear that they have read on the subject and want to communicate the reasons for their unbelief and that it is a valid choice. Atheism can take on a proselytizing function for some people, which is not unlike efforts to convert others to a religious belief system. Not unexpectedly, the more intelligent the unbeliever or the believer, the more elaborate and convincing the argument.

On the other end of the intelligence scale are developmentally delayed individuals who are not intellectually capable of meaningful participation in most religious activities. These individuals are not able to comprehend basic tenets of a religious creed, its logic, ethics, lifeway, or to interact with its supernatural beings in culturally prescribed ways. That is a great deal to learn, and again, this quantity of material (especially in non-literate societies with oral traditions) testifies to the need for a substantial amount of intelligence to remember and make sense of it all. It is important to remember that we refer to intelligence within an evolutionary context. There is a normal range of human intellect that is required for most types of religious thinking. If we assume that full intelligence comparable to today's modern humans has some relation to cerebellar re-organization, then we conclude that there is some minimum intelligence beneath which a species cannot engage in religious thinking. *Homo erectus*, for example, probably did not have the cerebellum (or other brain organs) of modern humans (indeed it is obvious from fossil skulls that the species did not), and this may be related to a lack of religious thinking. This is a preliminary contention on our part, but one based on an increasing number of findings from the modern sciences [32], [33].

We have often read authors (both those with belief and those with unbelief) who complain that religion removes the need to think and decide for oneself. Religion is sometimes seen as the "easy way," i.e., to go along with everyone else in accepting a theological interpretation. One simply does what religious doctrine dictates. However, religious thinking and participation are not quite that simple – a notion that is a surprise to many. Religious precepts take considerable thought to follow judiciously. True, religious beliefs can be "faked" but it takes a great deal of energy to lead a completely fake life. One corollary is true: If one finds that unbelief is the only tenable choice, that choice can have mild-to-severe social consequences. Choosing unbelief is a somewhat risky option, again underscoring the social foundation of religious capacity and its principal function in supporting the social group.

Building Blocks 3 and 4: Our Basic Ape Model; and, Better Vision and Hearing

Religious capacity is dependent upon the type of mammalian model we came from. The characteristics of our ancient ape ancestors were critical for what eventually emerges in our species as religious capacity. For our evolutionary model, we can see from primatology studies [14] that apes must have already been evolving in the directions that humans would assume in part, change in part, and use in

their own way. Therefore, it is useful to look at features of all apes, especially in modern apes. They are often called “relics” of a large and widespread populations of apes that were plentiful in the Miocene.

Apes are large, they develop slowly, they have big brains, and they are demonstrative, at times. They are also fundamentally social, they live in troops, and child-rearing is lengthy and intense, forming bonds that last for years. Apes also have no tail, which requires flexible and strong limbs and torso to stabilize movement and sitting. Those requirements are all retained by humans, who are flexible, demonstrative (at times), intelligent, and form deeply emotional bonds with other humans.

Humans who have no religious beliefs inherit all these traits, as do humans who have religious beliefs. We would not distinguish humans with unbelief on the basis of most of these features. We find few mechanisms that would cause religious capacity to fail to form, unless it is a fundamental failure in ontogeny in the development of sociality, intelligence, or ability to form emotional bonds with other people. Those incapacities could contribute to atheism or unbelief – or vary with it – because of the social and often emotional nature of so much, but not all, of religious experience.

We see the ape demonstrative tendency as fully congruent with many religious behaviors, especially for religious leaders but also for followers. Our improved vision and hearing make participation in religious activities substantially more intense. For example, human senses are fully at play in what we would term “numinous experience” that occurs so often in a variety of guises in most, if not all, religions. To the extent that improved senses heighten numinous experience, we conclude that heightened senses could support the emergence of religious capacity.

Building Blocks 5 and 6: Lengthened Developmental Trajectory; and Social Tolerance in Adults

During its lengthy evolutionary emergence, religious capacity depended on sociality in its basic form of living in troops (later groups or bands). It also depended on cooperative social activities of groups of adults and helping others with tasks of a physical or intellectual nature. Complex social participation by adults would not be possible with an aggressive orientation like that of modern chimpanzee. Religious activities often involve bringing other adults close, in order to prepare for or take part in rituals, teach lessons, and render assistance. Young anthropoid apes, like many immature mammals, socialize easily, so they are naturally accepted and included, but when adulthood is reached, this easiness is uncommon.

For our evolutionary model, changes had to occur to the more generally aggressive stance of adult great apes, or religious capacity could not have emerged. While ape juveniles play freely, adult apes tend not to socialize quite so much or so easily, with the exception of bonobos [20]. Our view is that their ancestors must have developed a parallel down-regulation of aggression, as in groups leading to humans. We hypothesize that a less aggressive adult style of interaction was achieved by extending juvenile socialization into adulthood, according to a “domestication” syndrome or suite of changes well outlined in other animals [20]. Modern apes show a variety of adult personalities and social styles, from the more aggressive chimpanzees to the more docile bonobos. Gorillas fall somewhere in between, although they can be aggressive when provoked, as can the rest, including the derivative human. Still, the level of cooperation achieved by adult humans is not seen in any other mammals.

Before the genera Homo and Pan diverged, other changes were happening to ancient Miocene apes that humans have inherited and accentuated. They involved a lengthening of the developmental life cycle, so that individuals matured more slowly, were dependent for longer as juveniles, and adults had greater longevity. The pattern is called, “secondary altriciality,” implying a secondary and a longer period of dependence (primary altriciality being at birth). Childhood lengthened and adolescence emerged, when before, there had simply been two age groups, the young and the adults. We hypothesize that, along with this lengthening maturation, some features of immature apes were retained into adulthood, especially the social tolerance of others and a behavior profile that encouraged it. Above, we even likened it to the profile for domesticated animals. In general, some ancient apes

became more tolerant as adults, with the possibilities opening up for cooperative group activities that emphasized interaction and heightened emotional experience. All of this would not be possible if some ancient apes had not evolved to mature more slowly with a changed temperament.

How does this emergence help us to examine atheism and unbelief in later humans? The answer may again come down to conformity and risk. We begin to see how the social exclusion of atheists may be something anticipated by them and, if necessary, tolerated. In other words, unbelief carries implications about sociality, tolerance, and cooperation. To choose disbelief requires some risk because of the possible exclusion of the atheist from some or all social roles and activities. We suggest that individuals with unbelief come to understand this risk and decide to either tolerate it or not. Indeed, people who profess unbelief in the modern world are taking a stand, which separates them from others who profess belief [30], [31]. To carve out a place for oneself that is separate runs contrary to many of the evolutionary social changes underlying religious capacity. Humans with unbelief naturally separate themselves from humans with religious capacity by their non-conformity alone.

Building Blocks 7 and 8: Further Upgrades in Intelligence to Manage Aggression; and Greater Sensitivity

When tolerance among anthropoid adults increases, so do the complexities of social life. This puts a premium on sensitivity, both a type of sensitivity that is connected to the network of emotion centers in the brain, and a type of sensitivity that is not, which is likened more to a type of awareness and readiness for action. Two types of sensitivity have already been connected to genomic segments, and scientists anticipate that there will be many more genes that affect sensitivity in the future [4], [1], [35], [44], [45].

Social sensitivity in humans is a complex characteristic whose genomic underpinnings are multiple, some known and some unknown. There is plenty of room for failure in the mechanisms and pathways that guide social life, as well as fractional activation at multiple levels and perhaps, additional loops and sequences that involve decision making and additional aspects of primate sociality.

Building Blocks 9 and 10: Biological Basis for Culture Emerges; and Then, Moral Capacity

In the ancient apes that eventually gave rise to the human and chimpanzee lines, around 8 to 10 million years ago, a cognitive capacity for culture arose. Today, we see culture only weakly in the chimpanzees and bonobos, but very strongly in humans. Culture is different from sociality, which characterizes all primates in many different configurations. Sociality is about group life – how it is configured, how dominance and nurturing are provided, and the group's functioning as a unit in provisioning themselves and defense from predators. It is said that social primate groups evolve as a group, and while individual members are the ones to convey the group's genetics, there is much truth to the notion that the group is the adaptive unit for evolution.

This changes things and sets the stage for the emergence of a trait that strengthens the group even more. Cultural capacity is biological, but culture itself is learned, passed on within the group, and because it is largely a cognitive trait, opens wide the possibility for cultural differences between groups, even if they sometimes interbreed. Culture can change over time, while new technologies are crafted, refined, and then discarded, and new customs emerge to define the group's identity. Its succession is not, itself, generally dependent on genetics, although we are learning more and more about culture-biology loops where the two adaptive systems affect each other.

Culture is learned by the young at a time when they are receptive. To the child, it appears natural, and they are unaware of learning much of it. It also takes attention by older individuals, purposeful teaching and learning, and therefore culture is highly dependent on the lengthened

developmental trajectory described above, and on early and long-lasting bonds between individual members of a troop – eventually, a human group, once the line to modern humans diverged around 6-7 million years ago. It was not until around 3 million years ago that the genus Homo emerged. By around 2 mya, *Homo erectus* emerged – fully bipedal, with a fully developed stone tool tradition that changed gradually. By about 1.5 mya, *Homo erectus* had learned to control fire, if the latest remains of fire making from Wonderwerk Cave in South Africa continue to hold [5].

It is in the species *Homo erectus* that we propose moral capacity first arose and was perpetuated culturally in a learning context we call "The Human Hearth" [32]. Fundamentally, moral capacity is a cognitive trait that is a biologically based facility for decision making based on neuronally encoded values [36]. Its phenotypic expression varies widely, but not as widely as often assumed, because there are broad cognitive similarities in the adjudication of different moral systems. For morality and most aspects of human life, culture defines what is real and not real, right and wrong, desirable and undesirable, for specific groups of humans. That gives several good clues about atheism, or the failure to form religious capacity in modern humans.

Religious thinking is a biological trait that uses cultural capacity to define cosmologies, to populate supernatural spaces, and importantly, to specify behavioral rules in the form of ethics. We have speculated that ethics derive, in general, from ideal characteristics of supernatural beings. Their qualities become guidelines for behavior to which humans can aspire. In many ways, the atheist could be defined as rejecting certain basic aspects of the cultural belief system held by people in his social group. The questions remain: Is this an intellectual choice, part of a biological trait that fails to form, or both?

We do not interpret cultural capacity, moral capacity, or the later, religious capacity, strictly as "adaptations" because there are strong indications that random genetic drift in small groups of apes and hominins may well have influenced the retainment of traits, in addition to natural selection. In particular, there is an absence of many "selective sweeps" in the human genome and the presence of much "neutral" material [21], [24]. Genetic drift is much stronger than natural selection in small groups, so the effects of natural selection were dampened. Genetic drift is a random, not a directional force, and was probably an important factor in the emergence of the special neurocognitive traits that make humans so special. We propose that cultural capacity probably arose in small groups of late apes or early humans, and remained because the groups were isolated, and natural selection was weak. The trait may have arisen more than once and may not have been immediately beneficial. It could even have been "slightly deleterious" – the term now used for traits that remained in the human genome and caused some damage, but not too much, or, in another environment, became beneficial.

Cultural capacity and the much later religious capacity emerged initially by chance, were retained, and later became useful to the social group. It is possible that religious capacity was retained because it had features that were physically and psychologically pleasing to humans. For early humans who were beginning to shoulder the burden of self-awareness, any altered state of consciousness might have had an appeal. By the time of *Homo erectus*, who hypothetically enjoyed a rudimentary morality and moral adjudication with the characteristics listed above, it is very likely that the human line was beginning to view its place in the universe (however that was conceived) and responsibilities to the social group were taking on moral dimensions. Recognizing these moral aspects of group behavior can be a heavy burden. Our contention is that this burden was, and remains, eased in *Homo sapiens* by religious capacity.

How can this inform our view of unbelief? It suggests that the atheist or agnostic has escaped the burdens of self-consciousness and found supports that can substitute for religious behavior and its expression. Another possibility is that the burden is borne entirely by science and the human intellect. This possibility forms an important part of the modern literature on atheism. What is noteworthy to us is that, for the most part, the argument is not adopted by many humans, at least yet. If religious capacity is the biological trait we identify, then it will come under enormous selection pressure in the modern

world, with its billions of humans. It is on large populations like the world (and soon off-worldly) human populations of today that natural selection will exert a newly strengthened influence. How will religious capacity fare after those pressures? Will religious capacity be genetically wiped out, or will it remain standing? Or, will it become something else?

5. Who Are the Humans in Whom Religious Capacity Fails to Form?

We have examined developments – biological innovations – that were foundational to religious capacity. We have speculated on reasons it may fail to form in some living humans. We conclude that religious capacity requires a basic sociality, but also special senses and sensitivities that enhance sociality and allow adults to approach each other and gather in groups for religious purposes. We have discovered that religious capacity entails good intelligence (from an evolutionary viewpoint), and that both sociality and intelligence are especially important for religious leaders. We conclude that any developmental problem that might impact social life and communication, or that severely limits intelligence, can well give rise to religious capacity's "failure to form."

We have also mentioned that religious and moral capacities involve intellectual decision-making capacities, and we have noted that many modern humans make a choice to adopt atheism or agnosticism for rational reasons. They know, understand, and appreciate social life, but are intellectually convinced that atheism is the right path. They can work well in groups, and sometimes stand out for their leadership. They tend to be socially aware, but simply choose another system of belief. Atheists who write of their "conversion" to unbelief, or atheists whose goal is to be "free" of religious thinking and experience, are generally in this group. Their belief system makes an intellectual statement and a life style choice. We note that agnostics and atheists sometimes appear in various states of indecision. This can be a very large number of people, which varies from society to society, and with stage in the human life cycle. Many young people question the religious and political beliefs of their parents. Is this a failure of religious capacity to form? We believe not. We see it as a stage in formation, where individuals consider the consequences of their involvement, and make a decision for belief or unbelief.

On the opposite end of the decision-making spectrum, we next mention a small percentage of humans who do not have moral capacity, and therefore religious capacity, at all. They have no "moral compass," and no decision-making apparatus that helps to define "right" from "wrong." These individuals can appear oddly out of tune to others when, for example, it comes to events that cause others to be squeamish, such as the killing of animals for no worthwhile purpose, or torture. They may find religious thinking beyond their ability, although they can appear quite intelligent in other ways, and live among others and even participate in religious activities. Still, they are often in an uneasy alliance, and others tend to sense their disjuncture. Individuals without a moral compass can be aware of it and good at hiding it. They tend not to have an interest in religious thought or experience, except for self-serving reasons.

We are left with questions about why some individuals choose to withdraw from religious participation, and why others remain fully committed to it. In the past century, in many modern and developing societies, we have seen the rise of a rationale for many females to withdraw from religious institutions, and it is seated in the traditional aspects of religious doctrine that they deem sexist, discriminatory, and even cruel. Yet, other women participate in religious activities and try to change rules from within. Socioeconomic factors can also be just as strong as gender in encouraging others to leave a religious institution or stay with it. And still, when all of the demographic factors are considered, we are not sure we understand why some humans have belief and others have unbelief.

Let us turn our questions about atheism and ask an obverse question. Instead of examining why atheists choose to relinquish religious experience, perhaps we should examine why others choose religion. Can a model of the evolution of religious capacity and knowledge of cognitive science and

genomics shed any light on why religious thinking and experience are sought? This approach may lead us closer to an understanding of why religious capacity can fail to form or not form fully. We are looking for some essential feature or features that believers want or need, which traces back to a critical evolutionary development on the human line. Our evolutionary "Building Blocks" explain a sequential foundation for religious capacity, and where it came from biologically. Now, we are required to explain how – or what – keeps it going. Why do more people *not* leave religion and opt for atheism, or at least agnosticism?

6. Origins of the Numinous and Its Modern Function in the Genus Homo

If we examine the features of ancient apes that had to change in order for religious capacity to flower millions of years later on the human line, we may find a clue that will help to explain humans' continued capacity and desire to experience religious thinking. It is not enough to call religious thinking a "biological trait," which it is. However, there is a choice concerning the expression of that cognitive trait, as there is with, for example, reading. This ability to decide can lead to some humans opting out, at whatever level of political and social development they exist, from tribal to modern nation-state. We search for something new and special that came with the human evolutionary line, which goes far in explaining why some humans steadfastly adhere to a life in which religious thinking is important. What keeps it going?

An analysis of atheism as a failure to form has unexpectedly led to conclusions about religious capacity's fundamental nature, why it has been successful in supporting the social group, and why it has become so bound up with social control. When we listed the evolutionary upgrades that occurred in some groups of late Miocene apes, we included improved senses, especially in vision and hearing. We also noted how these senses came to heighten the experience of modern humans in religious activities and to make their experience more intense, and in some way, more satisfying. We also described a tendency for a demonstrative quality in ape behavior, from time to time, and found it fully congruent with many religious and artistic behaviors in their descendants. Religious activities involving the senses range from the very active, loud, and dramatic, to the quiet, thoughtful, and subtle. To the extent that improved senses today heighten the experience of religious participation and lead to what some have labeled "a numinous state," we conclude that the heightened senses emerging in some ancient apes could have supported the evolution of religious capacity, and perhaps, helped to keep it going in the lives of humans.

The observation of a connection between types of senses that were upgraded on the human line, and continued capacity and desire for religious thinking may be a new type of observation, but the observation that the numinous is important is surely not new. Still, our goal is to understand religious capacity's failure to form, and in biologically determined senses we may find a clue. Where these senses are not present, or not finely tuned, their absence could help to explain atheism. Those who choose atheism may not achieve a sensory "high" from religion, and therefore choose unbelief because of the long-researched "costs" associated with religious participation. When questioned, the humans who believe in the supernatural repeatedly refer to experiencing something like "numinous experience" as one of religion's main attractions. We add one additional and very important factor. Individuals on the human line, at least from the time of *Homo erectus* [13], had a daily model of the numinous to which they could point – dreaming.

We have crafted our contention very carefully, because for some scholars, the "numinous" is non-sensory, non-rational, and not detectible by science, i.e., it is a "non-sensory feeling" [28]. The latter could be called an oxymoron, and indeed we believe it is. While some altered mental states may seem divorced from sensory experience (like dreaming, hallucinations, and drug states, to name a few), we understand the human body to be the interface with the outside world. Sensory experiences based in the environment are coded in the human nervous system, even if they come to be interpreted differently

later. We do not believe that the notion of "numinous experience" needs to be, or indeed can be, separated from the human senses. Therefore, we look to human senses as a model for numinous experience. Often, sensory terms are all that humans have to describe the numinous.

The notion that experiencing the numinous is separate, outside the self, and that the individual experiences it as "ganz Andere" [German, completely different] or "wholly other" is a very appealing idea that, in its error, manages to capture much of the appeal of religious experience: It takes the person somewhere that he or she usually does not go. In understanding the appeal of the numinous, we come to see that it is seated in normal human sensations that are derived from the experiences of the human body [27], but that the numinous often comes to be interpreted as supernatural. Why would the experience of a mental state called "numinous" be so widely appealing, become so culturally meaningful, and why is it connected so easily to the supernatural – worldwide?

For the atheist, there must be a fault in the following progression. Either the atheist does not experience the numinous, does not want to, finds it too frightening, does not find it culturally meaningful, or does not connect it to the supernatural. Furthermore, the atheist surely has difficulty with religion's unfailing support for the social group and especially, its connections to social control. Let us take a closer look at the concept of the numinous, and how it strikes scholars and the devout, alike.

Rudolf Otto, who helped to popularize the concept of the "numinous" [28], believes that it cannot be defined in terms of other experiences. He finds that it is not perceptible by human reason, and it is a mystery both terrifying and fascinating [*mysterium tremendum et fascinans*] [2]. In *The Idea of the Holy*, Otto writes:

> The feeling of it may at times come sweeping like a gentle tide pervading the mind with a tranquil mood of deepest worship. It may pass over into a more set and lasting attitude of the soul, continuing, as it were, thrillingly vibrant and resonant, until at last it dies away and the soul resumes its "profane," non-religious mood of everyday experience... It has its crude, barbaric antecedents and early manifestations, and again it may be developed into something beautiful and pure and glorious. It may become the hushed, trembling, and speechless humility of the creature in the presence of – whom or what? In the presence of that which is a Mystery inexpressible and above all creatures [2].

From an evolutionary perspective, the following words are particularly important: "It has its crude, barbaric antecedents and early manifestations…" These are the origins of the numinous in upgraded human senses. We ask: When and how did the numinous emerge as a mental state that members of our genus Homo could enjoy and in which they could find solace and relief? It may have begun at the time of *Homo erectus*, in a context we call "The Human Hearth," where a *Homo erectus* band met around a campfire to recount stories (in a language that was not yet fully grammatical according to some, and fully grammatical, according to others). They discovered the right and wrong of the day, and first conceived of a supernatural spirit [32-33]. Is that when the numinous began to be appreciated? Perhaps, or perhaps it had to wait for the larger capacity of the *Homo sapiens* brain, especially portions of the parietal lobes and the precuneus, whose partial de-activation is known to cause dream-like states. According to research, the modern precuneus could be capable of manipulating counterfactual places and beings. Religious thinking could not exist without the expanded precuneus of our species. Still, to follow the logic of evolutionary science, it almost surely began with antecedents we can begin to identify now.

7. Conclusions

We have worked our way through the biological origins of religious capacity, and we have searched for antecedents of present-day types of thought and activities. We first asked how religious capacity formed, and then speculated on how it could "fail to form" and lead to atheism or unbelief. We identified some types of humans in whom religious capacity does not develop, take hold, and emerge according to cultural motifs of the person's society. However, we still could not identify, at first, the feature that "failed to form." Therefore, instead of asking why and how religious capacity could fail to form, we asked the obverse: What keeps it going? Why do humans not all simply relinquish religion and all its strictures, and take up a life of unbelief? In the progression we sketched earlier, we find all types of evolutionary innovations that led to religious capacity, from a downregulation of aggression to an upgrade in social sensitivity. The changes in the senses were simply noted as important because they help to make religious group activities so meaningful, with songs, chants, lighting, shadows, and the drama of religious ritual. We searched and found the seemingly, only vaguely important upgrades in the senses. That suggested some change in the senses that might give humans motivation to come back to religious thinking, time and again. It had to be something that relied on evolutionary changes but gave them motivation, now and into the future.

We had to go further. In our focus here on the numinous, we slip from direct sensory experience to sensory experience that is not, ostensibly, of this world. That is the "theory" embedded in the "theology." The fact that it is indeed of this world can be logically set aside by believers because of its appeal. In the course of religious worship, the worshippers do not care. Religious doctrine and ritual make the other-worldly, real. Recall, culture defines what is real and not real, and it accomplishes this through religion and other institutions. The numinous combines known sensory experiences into a jumble of vague feelings that lose their comparison with mundane human life. The numinous provides an altered state of consciousness that can be easily sought, obtained, and left behind when a religious activity ends. There is no need for drugs, although drugs are often used in religious rituals. There is no need for extremity or pain to achieve a hallowed state. The numinous is its own, low-level, sensory "high" and it forms an important part of the basis for why many humans participate in prayer, offerings, services, music, singing and chanting.

If desired, atheists and agnostics can find the numinous in nature, as did Ursula Goodenough in *The Sacred Depths of Nature* [18]. On the other end of the intensity scale, Alvin Toffler, in *Future Shock* [46], foresees an entire industry of "Experience Makers," while the manufacturing and service industries fade away. "Experience" will be a commodity like all others – planned, exquisitely packaged and delivered, and often completely simulated. Why would this be the industry Toffler foresaw – other than the fact that even in 1970, it had already begun to form?

What do we have now that meets those needs? We have art, music, science, and we have religion. In mankind's long progression from his origins among the ancient apes of Africa, he has sought release, amusement, delight, and escape from the self-consciousness that evolution handed him. It is a heavy burden, to see oneself on a timeline, to know of one's ultimate demise, to realize that mistakes can be made that hurt others, and to willingly accept the ultimate control of the social group. All of this is a burden that needs healing. We as a species were dealt a heavy blow. We know right from wrong. We live self-consciously and tentatively, and we relish the experiences that can shoulder part of that burden just a little bit of our time on Earth.

8. Epilogue

Not surprisingly, the two authors have different perspectives in light of the analysis in this paper. One, an anthropologist, is probably closest to an agnostic, although she has experienced the numinous in religious services. The other, an astronomer but also Catholic priest, sees "crossing over" to a belief in

the numinous as “a gift.” In spite of these two very different perspectives, it is interesting that the analysis stands.

References

1. Acevedo, B. P., Aron, E. N., Aron, A., Sangster, M.-D. Sangster, Collins, N., Brown, L. L. The Highly Sensitive Brain: An FMRI Study of Sensory Processing Sensitivity and Response to Others’ Emotions, *Brain and Behavior* 4, 2014, pp. 580-594.
2. Alles, G. D. Ed. *Autobiographical and Social Essays by Rudolf Otto*, De Gruyter Mouton, 1996.
3. Barth, F. *Nomads of South Persia: The Baseri Tribe of the Khamseh Confederacy*, Little, Brown, 1961.
4. Beevers, C. G., Ellis, A. J., Wells, T. T., McGeary, J. E. Serotonin Transporter Gene Promoter Region Polymorphism and Selective Processing of Emotional Images, *Biological Psychology* 83, 2009, pp. 260-265.
5. Berna, F., Goldberg, P., Horwitz, L. K., Brink, J., Holt, S., Bamford, M., Chazan, M. Microstratigraphic Evidence of In Situ Fire in the Acheulean Strata of Wonderwerk Cave, Northern Cape Province, South Africa, *Proceedings of the National Academy of Sciences* 109, 2012, pp. E1215–20.
6. Bruner, E., Iriki, A. Extending Mind, Visuospatial Integration, and the Evolution of the Parietal Lobes in the Human Genus, *Quaternary International* 405, 2016, pp. 98-110.
7. Bruner, E., Pearson, O. Neurocranial Evolution in Modern Humans: The Case of Jebel Irhoud 1, *Anthropological Sciences* 121, 2013, pp. 31-41.
8. Bruner, E., Preuss, T. M., Chen, X., Rilling, J. K. Evidence for Expansion of the Precuneus in Human Evolution, *Brain Structure and Function* 222, 2017, pp. 1053-1060.
9. Bruner, E., Spinapolice, E .E., Burke, A., Overmann, K. A. Visuospatial Integration: Paleoanthropological and Archaeological Perspectives, In *The Evolution of Primate Social Cognition*, Springer, 2018, pp. 299-326.
10. Cavanna, A. E., Trimble, M. R. The Precuneus: A Review of Its Functional Anatomy and Behavioural Correlates, *Brain* 129, 2006, pp. 564-583.
11. Cavanna, A. E. The Precuneus and Consciousness, *CNS Spectrums* 12, 2007, pp. 545-552.
12. Coolidge, F. L. Exaptation of the Parietal Lobes in Homo sapiens, *Journal of Anthropological Sciences* 92, 2014, pp. 295-298.
13. Coolidge, F. L., Wynn, T. *The Rise of Homo sapiens; The Evolution of Modern Thinking*, Wiley-Blackwell, 2009.
14. De Waal, F. *Primates and Philosophers: How Morality Evolved*, Princeton University Press, 2009.
15. Dixon, M. L., De La Vega, A., Mills, C., Andrews-Hanna, J., Spreng, R. N., Cole, M. W., Christoff, K. Heterogeneity within the Frontoparietal Control Network and Its Relationship to the Default and Dorsal Attention Networks, *Proceedings of the National Academy of Sciences*, 201715766, 2018. Published online ahead of print. https://doi.org/10.1073/pnas.1715766115
16. Gazzaniga, M. S. The Interpreter Within: The Glue of Conscious Experience. Dana Foundation web site, Gazzanig’s blog, 1999.
http://www.dana.org/Cerebrum/Default.aspx?id=39343#sthash.I7zCiFeL.dpuf
17. Gómez-Robles, A., Sherwood, C. C. Human Brain Evolution; How the Increase of Brain Plasticity Made Us a Cultural Species, *MÉTODE Science Studies Journal* 7, 2016. [no page numbers]
https://doi.org/10.7203/metode.7.7602
18. Goodenough, U. *The Sacred Depths of Nature*, Oxford University Press, 1998.
19. Gunz, P., Tilot, A. K., Wittfeld, K., Teumer, A., Shapland, C. Y. Shapland, Van Erp, T. G. M., Dannemann, M., Vernot, B., Neubauer, S., Guadalupe, T., Fernandez, G., Brunner, H. G., Enard, W., Fallon, J., Hosten, N., Volker, U. Profico, A., Di Vincenzo, F., Manzi, G., Kelso, J., St. Pourcain, B.,

Hublin, J.-J., Franke, B., Pääbo, S., Macciardi, F., Grabe, H. J, Fisher, S. Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity, *Current Biology* 29, 2019, pp. 120–127.
20. Hare, B., Wobber, V., Wrangham, R. The Self-Domestication Hypothesis: Evolution of Bonobo Psychology Is Due to Selection against Aggression, *Animal Behaviour* 83, 2012, pp. 573–85.
21. Harris, E. E. *Ancestors in Our Genome: The New Science of Human Evolution*, Oxford University Press, 2015.
22. Hicks, J. M., Coolidge, F. L. On the Role of Precuneal Expansion in the Evolution of Cognition, Published online, University of Colorado, Center for Cognitive Archaeology, Colorado Springs, 2016.
23. Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*, *Nature* 546, 2017, pp. 289-292.
24. Lachance, J., Tishkoff, S. A. Population Genomics of Human Adaptation. *Annual Review of Ecology, Evolution, and Systematics* 44, 2013, pp. 123–43.
25. Lang, S. *Men as Women, Women as Men: Changing Gender in Native American Cultures*, University of Texas Press, 1998.
26. Marx, K. *A Contribution to the Critique of Hegel's Philosophy of Right.* Press Syndicate of the University of Cambridge, 1967 [orig. 1843].
27. Newberg, A. B. *Principles of Neurotheology*, Ashgate Publishing Limited, 2010.
28. Otto, R. *The Idea of the Holy*, Oxford University Press, 1923.
29. Pääbo, S. *Neanderthal Man: In Search of Lost Genomes*, NY: Basic Books, 2015.
30. Pew Research Center. The Religious Landscape Studies, online, 2012.
31. _____. Why Americans Go (and Don't Go) to Religious Services, online, 2018.
32. Rappaport, M. B., Corbally, C. The Human Hearth and the Dawn of Morality, *Zygon: Journal of Religion and Science* 51, 2016, pp. 835-866.
33. _____. Human Phenotypic Morality and the Biological Basis for Knowing Good, *Zygon: Journal of Religion and Science* 52, 2017, pp. 822-846.
34. _____. Evolution of Religious Capacity in Genus Homo: Origins and Building Blocks, *Zygon: Journal of Religion and Science* 53(1), 2018a, pp. 123-158.
35. _____. Evolution of Religious Capacity in the Genus Homo: Cognitive Time Sequence, *Zygon: Journal of Religion and Science* 53, 2018b, pp. 159-197.
36. _____. Evolution of Religious Capacity in the Genus Homo: Trait Complexity in Action through Compassion, *Zygon: Journal of Religion and Science* 53, 2018c, pp. 198-239.
37. _____. In press. Importance of the Activation and Deactivation of the Precuneus in Human Theological Thinking and Experience of Immanence and Transcendence, *Studies in Science and Theology*, ESSSAT, 2019.
38. _____. In press. Cultural Neural Reuse, Re-deployed Brain Networks, and Homologous Cultural Patterns of Compassion, *Proceedings, Illuminating Biological Systems from a Network Perspective*, University of Namur Press, 2019.
39. Rendu, W., Beauval, C., Crevecoeur, I., Bayle, P., Balzeau, A., Bismuth, T., Bourguignon, L. et al. Evidence Supporting an Intentional Neandertal Burial at La Chapelle-aux-Saints, *Proceedings of the National Academy of Sciences* 111, 2014, pp. 81–86.
40. Sack, A. T. Parietal Cortex and Spatial Cognition. *Behavioural Brain Research*, 2009, pp. 153-61.
41. Smaers, J. B., Turner, A. H., Gómez-Robles, A., Sherwood, C. C. A Cerebellar Substrate for Cognition Evolved Multiple Times Independently in Mammals, *eLife Sciences* 7, 2018.
42. Spreng, R. N., Mar, R. A., Kim, A. S. N. The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind and the Default Mode, *Journal of Cognitive Neuroscience* 21, 2008, pp. 489-510.
43. Tanabe, H .C., Kubo, D., Hasegawa, K., Kochiyama, T., Kondo, O. *Cerebellum: Anatomy, Physiology, Function, and Evolution. Digital Endocasts, Replacement of Neaderthals by Modern Humans Series*, Eds. Emiliano Bruner et al., Springer Japan, 2018, pp. 275-289.

44. Todd, R. M., Ehlers, M. R., Müller, D. J., Robertson, A., Palombo, D. J., Freeman, N., Levine, B., Anderson, A. K. Neurogenetic Variations in Norepinephrine Availability Enhance Perceptual Vividness, *Journal of Neuroscience 35*, 2015, pp. 6506-6516.
45. Todd, R. M., Müller, D. J., Palombo, D. J., Robertson, A., Eaton, T., Freeman, N., Levine, B. Anderson, A. K. Deletion Variant in the ADRA2B Gene Increases Coupling between Emotional Responses at Encoding and Later Retrieval of Emotional Memories, *Neurobiology of Learning and Memory* 112, 2014, pp. 222-229.
46. Toffler, A. *Future Shock*, A Bantam Book, 1970.
47. Wynn, T., Coolidge, F. L. *How to Think Like a Neandertal*, Oxford University Press, 2012.
48. Zhang, S., Li, C. R. Functional Connectivity Mapping of the Human Precuneus by Resting State fMRI, *Neuroimage* 59, 2012, pp. 3548-3562.

Notes

1. We specify “adult” behavior and cognition because many children cannot always distinguish between fantasy and reality. Since religious experience sometimes involves altered states of consciousness that mix real and unreal, conscious and unconscious, dream and wakefulness, we believe it is reasonable to specify that only adult humans experience “religious thinking”. Of course, children can begin to learn about religion early, and practice religious activities, but religious thinking is a domain primarily for adults.

Studia Humana
Volume 8:1 (2019), pp. 67—74
DOI: 10.2478/sh-2019-0004

Dealing with Free Will in Contemporary Theology: is It Still a Question?

Lluis Oviedo

Antonianum University,
Rome, Italy

e-mail: loviedo@antonianum.eu

Abstract:
Free will is a very hot issue in several theoretical settings, but less in theology, or at least not as much as use to be in former times, when the discussions on sinfulness, grace and freedom were igniting a long season of controversies, especially in the Reformation time. Even in ecumenical dialogue apparently free will does not play a great role, since the reached consensus seems quite peaceful and agreement dominates over discussion. However, some theological insights, especially Karl Rahner reflections, are still worthy to consider and possibly theological anthropology should pay more attention to the current debate and its consequences for the way we understand human nature and its relationship with God.
Keywords: Karl Rahner, Christian anthropology, Imago Dei, sin, grace.

1. Preliminary Layout

Theological treatment of free will poses a good 'case of study' to test Christian anthropology and its adequacy to new cultural and philosophical settings. Several new theological issues related to freedom in human beings arise in the new context, when comparing with traditional views. That topic is involved in the three main principles that theology trying to understand human person has always claimed: the *Imago Dei* human attribute, which includes freedom as a postulate or condition to proper talk about similarity to God; the failure state expressed as sinfulness, which has been viewed in terms of freedom trimming or limited; and the effects of grace, including among them – with some nuances – a restored full capacity for freedom, liberation from sin's bondage.

Traditional and confessional positions have enriched a long discussion trying to better assess the extent that free will can reach, or its limits and boundaries, due to sin and other human limits. However, in some way, old confessional discussions seem to be overcome by new awareness and the revision of early held positions concerning the corrupting effect of sin and the re-generating consequence of grace. This is a quite realistic approach which has always moved between the empirical observation regarding human behaviour, and the speculative reflection inside a revealed framework. Theology describing or discerning free will has never been only 'speculative' and *a priory*; often the stated positions reflected internal struggles, or a way to observe human nature around, with its many trials, and its dark side too. However, the Biblical text has functioned as a

framework that provided a way to interpret and understand what was being lived and what could be perceived and wished as salvation. To some extent, the long history of controversies around free will in Christian theology reflects developments in the environment and broader anthropological positions in their time *Geist*, with its openness and despair, with its more or less optimistic feelings; modelled on historical circumstances of achievement and frailty.

Now the question that still looms is whether the discussion we drag for centuries is still a 'theological question' that deserves to pay much more attention, or this is a question around a subject that has already known every possible answer, and reflects an exhaustion state. Perhaps nowadays it might be better to leave it to philosophers and to their distinctions, somewhat alien to theologians, tired after a long discussion period that appears to many of little use today. Possibly the topic has reached some maturity and most theologians are convinced about a standard position that satisfies every side. Furthermore, the issue has been displaced by other worries and more urgent challenges linked to Christian faith and its survival in very secularized societies.

'Freedom' in no longer a theological hot topic, at least in Catholic theology, where it is hard to find new contributions – besides the handbooks covering Christian anthropology and its historical process – and able to deploy models that, at least, would be able to engage with recent developments in philosophy and scientific study of human nature. However, in my opinion, this is something that still needs to be assumed and tackled. The ideal of freedom has given place in Catholic theology, since the seventies, to the so called *Liberation theology*, reflecting more practical concerns about the huge contrast between what was stated by the standard theology of freedom and salvation, and the reality in which entire populations were living, in conditions which did not allow to enjoy the demanding standards linked to Christian models of free will. These practical issues can be traced more generally to moral theology and its concerns regarding responsibility and accountability depending on how much free will can be recognized.

In the present reflection, my aim will be twofold. First, I will engage in a dialogue with Karl Rahner's attempt – one of the last and more originals – to render a completely updated theological account on human freedom, showing its full theological character and its limits and paradoxes. And second, I will try to figure out what could become a theological agenda for dealing with free will, after recent developments, both in the philosophical discussion, and in scientific research, both designing a completely new context for theological reflection. Possibly a third point could be offered, connecting with the former points: the centrality of freedom for theology requires to connect it with love's experience and commandment. This is however a point that needs to be developed in a different study.

2. A 'Modern' Theology of Freedom

A theology of free will – very controversial – has always existed, since the Patristic times, through medieval disputes until Reformation times. Describing the extent of human freedom has been central in the anthropological reflection moved by great authors. A distinction that signed the School theology in Middle Ages has been the greater or lesser space recognized to human freedom, as a result of two different broad systems: one based on the 'universals' theory, held by Thomas Aquinas, and limiting the reach of free will into a great created and harmonious plan; and a system giving more space to contingency and free process at various levels, which was characteristic of Duns Scotus and other Franciscans, and giving place to a more unpredictable, open and free world. The issue of freedom was hence deeply entrenched in big cosmological and epistemological views, it was a substantial part of the wide world view held by different theological schools.

Some attempts can be found in modern times to update theological motives, as for instance in Kierkegaard radical treatment concerning decision. In Catholic terms possibly the most interesting and updated attempt – for his own time – to develop a theology of freedom was moved by Karl Rahner, mostly in an article published in 1965 [9]. The title is very explicit: *Theology of freedom*, and it deals with that issue inside his own theoretical framework: transcendental neo-Kantian anthropology. The question now is what can we still learn today from that attempt and

which points require some updating, complement or even deep revision. The points I will focus more are the following: the Christian historical vindication of freedom; the very theological character of free will; the paradox of theologically understood freedom; and the relational character of freedom in that context.

a. Christian Radicalization of the Concept of Free Will

The paper starts with a strong statement, almost apologetic:

> … the real freedom of choice as such – i.e. the freedom which consists not only in the fact that man cannot be forced from without but also in that a free decision about himself is demanded from him which, therefore, is rather a demand and a task than freedom – can alone already be seen quite clearly in Christianity [9, p. 179].

In other words, only inside Christian faith human beings become fully responsible before the eternal God's love and demanding the highest responsibility. The revelation in Christ assumes a funding character for human freedom, which becomes exalted to the highest imaginable level, as far as it becomes freedom from God and towards God.

It is interesting, nevertheless, to consider other texts regarding the Christian idea of freedom and its relationship with philosophy. For instance, a brief Encyclopedia entry from 1975 states:

> For a systematic theology of freedom which will go beyond the framework of the post-Tridentine systems of grace and free will, only preliminary suggestions are provided by modern philosophy. The basic principles of a theological anthropology will point the way to a deeper grasp of freedom [8].

Previously Rahner reminds that "The theological notion of freedom was carried on from the start in a dialogue with the philosophical notion of freedom throughout its history" [8, p. 544]. Modern times have developed a new stage to the analysis of free will, but the author complains that the modern debate has been scarcely received in theology. Rahner's aim is to fill that void and to provide a fruitful theological elaboration of the available ideas elaborated by modern thinkers. The critical point seems to be that Rahner uses the new frame to better appreciate the value and meaning of the traditional Christian view. He seems to be trustful to the enunciated principle: a Christian view on freedom was closely entrenched with the philosophical reflection – at least until modern times; such dependency has been unjustly broken and needs to be restored as a condition to better appreciate the deep meaning of Christian freedom, and to exalt its value. It appears that only inside the reference to the modern secular philosophy of freedom, we can recognize the theological meaning of the Christian contribution.

A first consideration comes to mind: Rahner's thought can be placed in good company with contemporary authors who vindicate the central role played by Christian faith in configuring modern mentality and values. Very recent titles like: Larry Siedentop, *Inventing the individual: The origins of Western liberalism* (2015), and Nick Spencer, *The Evolution of the West: How Christianity has shaped our values* (2016), witness to an historiographic trend that moves in a similar direction: to vindicate the necessary role played by Christian faith in the modern development of liberal ideas, which, at the same time, are founded on a deep trust on human free capacity to decide and to find the most convenient course of action, at the individual and the social level. The related question concerning how far can modern liberal societies go when such original impulse fades away is not clear, even if very pessimistic forecasts are available, especially at the hands of *Radical Orthodoxy* theologians and their program of deconstruct modern and secular ideas [4], [5].[1]

A second consideration comes to mind when the temporal distance is taken into account, regarding Rahner diagnostic. After several decades from that subtle criticism, still similar concerns

arise: philosophical reflection has moved forward and since the sixties an abundant philosophical corpus has dealt with free will in almost all possible ways. Once more, it seems that the theological appreciation of freedom should recover from the delay that Rahner perceived in his own time, to accomplish a sort of historical destiny: freedom can be theologically understood only in close dialogue with one's own time philosophical reflection. This is nowadays an unfinished theological business, except that we admit that such a task has been accomplished – and quite decently – by philosophical theology, the discipline that would have assumed that reflective role, after some theological neglect, disinterest or even fatigue after a long and seemingly fruitless discussion several centuries earlier.

b. Free Will as a Fully Theological Category

In Rahner's analysis, human freedom is seen in pure theological terms, to the extent that radical freedom can be only understood in reference to the divine mystery, and that freedom reveals the greatness of divine's gift to humans. An almost 'Anselmian argument' version emerges here: the greatest that can be conceived in humans is the highest degree of freedom, but that possibility can only be reached if it is related to God and sustained by Him. Rahner program seems oriented towards recovering the lost theological dimension that was missed – or perhaps secularized – in the modern approach to free will, and this can only be reached when such an experience is explicitly linked to the divine being.

This is an old and often told story: modern thinking was suspicious regarding divine dependency, as a condition which would result in a trimming of human freedom. As Kant did stress very explicitly, modernity is vindicated as autonomy from external authorities, and God was surely in that list. In other words, the divine was thought ever since and at different modernity stages in terms of heteronomy, dependence, alienation, and a limiting power or presence. Rahner manoeuvres – as has always done – in the opposite sense: where others have seen God as a competing power, Rahner stresses his necessity to found and render possible human freedom; where others see in God a limiting presence, Rahner finds it as an instance of empowerment – applying a contemporary terminology! Christian faith follows in Rahner's version always a similar pattern: it becomes the best way to encourage human awareness and to render possible what otherwise would be hardly conceivable. In short, God becomes the condition of possibility to human freedom; and the freedom we can experience is always placed in a horizon of divine gift.

Some consideration comes to mind in this case as well. To some extent what is here on play is the hypothetical secularization of the Christian understanding – and foundation – of free will. As Karl Löwith has pointed at the same time when Rahner was developing his analysis, modernity could be understood as an exercise in usurpation and reconversion of Christian topics to become fully secularized and placed in a different context, to be reused and serve other interests, once they have been deprived from any theological reference [3]. Rahner efforts seem to point to a recovery of modern topics and ideals inside a Christian matrix, to show that they can work very well in the religious context, perhaps after a convenient lifting and updating to adapt them to the modern times and mentality. Christian faith in God's presence at historical and anthropological levels renders the modern project theologically legitimate and explained.

At this point it becomes unavoidable the reference to Charles Taylor and his attempt in *A Secular Age* [10] to correct a trajectory in Christian praxis – especially in Catholic style and magisterium – to come to terms with that modern development, and to adapt to the new situation signed by the expressivist turn, as exposed in the book *Sources of the Self* [11]. The question now is to what extent Rahner's endeavour was successful in his time, and how this exercise at 're-theologizing' a topic that was fully transferred to the secular realm could be re-assumed in the Christian mind. Our philosophical time would feel possibly uncomfortable with the transcendental categories that Rahner applied, and which were quite usual in the German context of those years, and perhaps ignored in English speaking areas. However, the challenge he was able to address is still looming for us, two generations later.

c. Paradoxes Arising From Freedom Theology

The mentioned article reports about a big paradox at the centre of Rahner's transcendental treatment of that human trait: God is at the same time the foundation or condition of possibility of human freedom, but, nevertheless, humans can use that faculty to deny their own source or foundation: "that freedom, however, is freedom *vis-à-vis* its all-supporting ground itself, that in other words it can culpably deny the very condition of its own possibility in an act which necessarily reaffirms this condition is the extreme statement about the nature of created freedom" [9, p. 181].

"God is affirmed and denied at the same time" [9, p. 181]. Such paradox is well described by the own theologian, and by others. The topic has deserved even a monographic research [6].

In principle, Rahner can feel close to Kant and other philosophers exposing the 'freedom antinomies', a classic of modern thought. However, in this version the paradox appears as fully theological, or as another 'theological paradox'. Critical voices can point to the flaws in the transcendental pattern serving Rahner's program, and which would be guilty of a form of self-referential paradox. The interesting thing is how Rahner manages to address the challenge, and to point to a radical level in which freedom is at the end the possibility of a self-negation (not in 'kenotic sense', of course!), at the time that negates its own foundation. To some extent such decision, possible, brings to light the definitive character attributed to freedom and the tragic consequence lurking in such decision.

In that argument arises something quite intriguing, since the free decision towards God, the 'ground' of free decision, entails a very self-destructive consequence, something perhaps too costly when conceiving freedom's foundation in that high theologically loaded view; a complete failure in self-understanding, a deep alienation appear as a normal consequence [9, p. 185]. Perhaps a less theologically intense concept would carry less severe consequences at the anthropological level; the price can be seen as too high for assuming that theological foundation. This point is probably linked to other modern understanding of free will as a radical decision which would endow with meaning one's own life. Surely Kierkegaard comes to mind, as does the XX century existentialism exploiting similar views about the radical character of life bounding decisions.

The question now is how much dated is that view, and whether covering human freedom with that radical theistic meaning still makes sense, when the cultural environment has changed so much, existential concerns have been downplayed and a theological-radical view of freedom appears to today sensitivity as quite far from what is felt and lived in broad cultural settings. Some normative issue is at stake in this case, and a question opens, in the sense that possibly a correlation can be found between the complete secularization of freedom and its devaluation and even banalization in current cultural terms. Are we again before a new version of the modernity malaise?

d. The Relational Character of Freedom

In Rahner's analysis, freedom clearly serves the cause of God's love: it is freedom to love God, since this is the 'fundamental human act', the one projecting sense on every aspect of human life, redeeming it from all its 'darkest hours', and helping to cope with risks and sin. That love is the 'human integrating principle'. However new issues arise inside this attempt at conceiving God's love and its foundational character, its 'athematic' or 'transcendental' condition. It seems that such a condition could prevent a real experience of love: you can hardly love what is constituting yourself, comprising human freedom. Here Rahner resorts to a distinction to avoid that new difficulty: between that divine previous and constitutive presence, and the one which can be thematised or expressed in 'categorial' terms. But the real answer lies in the mediation of love as neighbour's love, in whom the original relational character of freedom can be fully expressed and lived.

Again some suspicion arises in that schema, based on a strong theological description of human freedom, with all its attached conditions. The problem can stem from a perspective that reflects modern anthropologies, built on the individual and its transcendental constitution, instead of

building from a relational schema that privileges alterity. Possibly Rahner's view is still too much self-centred, even if the human proper foundation is given from outside, and relates to God as a source of being and freedom; the alterity is expressed in terms that are still too much referred to one's self, and less to an external and original experience of calling and relatedness, as has been stressed by many authors moving from the perceived Enlightenment individualism. This is the danger of the transcendental categories that Rahner relies on: that at the end the modern self-sustaining individual and the one intimately and secretly founded in the Divine presence become undistinguishable. Rahner seems to work in a time not yet deeply touched by the alternative anthropology built on the priority recognized to other's presence or the external input they provide, the one that has been championed by the Jewish philosopher Emmanuel Levinas, among many others. From such an anthropology a different theology of freedom would be required and could be built. Indeed the relational character of freedom in Rahner's version appears as not relational enough.

3. What Needs to Be Assumed in a More Updated Theology of Freedom

Engaging with Rahner's theological treatment of free will offers a good opportunity to review an effort made half a century ago, fully committed to a strong philosophical strand of his time and to learn from that endeavour to engage in our days with that thorny issue. Possibly the task remains open and invites to follow those footsteps when free will is considered still a theological question, and not something already settled and a topic too much discussed in past centuries and now perceived as tiresome.

I will propose three strategies or moments as a program aimed at updating freedom's theological approach: learning from the current philosophical debate; assuming a more explicit 'empirical stance'; and considering freedom inside the believing process as a general framework.

a. Learning From the Current Philosophical Debate

Anybody acquainted with the contemporary treatment of free will can recognize the spectre stretching between the extreme positions of 'determinists' and 'libertarians' and the somewhat in the middle 'compatibilists'. We count with excellent descriptions of such rich spectrum, as for instance the excellent systematic review provided recently by our colleague Aku Visala [13]. The least we can say is that the described panorama speaks for an unavoidable pluralism and an unsettled discussion in which the different parties have good arguments to support their own positions. Probably this is the current situation and there is no reason to expect that things will change in the near future.

The question now is what can theology learn from that state of things, provided that we still admit that theology can take advantage from a dialogue with the philosophy produced in our days, instead of keeping more self-referential and dealing with its own tradition and former ideas, what can be appreciated as 'classic'. Possibly theological development can feel some familiar sensation: we have been already there, could say the theologian used to a hermeneutical and historical analysis. Theological controversies, at least since the time of St. Augustin and his struggle with Pelagius, and those associated to the different Reformation versions, resulted again in an unsurmountable pluralism, this time reflected between confessional lines. Two ideas come to mind: the first is that perhaps we need to recognize that both, in the Christian and the secular philosophical realm, dealing with freedom means to struggle with too many antinomies, paradoxes and even contradictions, and hence possibly Rahner was right when describing "Freedom as a mystery" [9, p. 190], at least in the sense of posing many challenges to a reason trying to fix and to better describe it. Perhaps if we could better know and determine freedom, we would, as a result, become less free, in the same way that evil is a mystery: if we would better know evil it would become less bad. A better comparison arises with the mystery of grace, whose complete knowledge would render it less 'gracious' and effective. Since freedom is linked in Christian theology with all these categories: sin

or evil, love and grace, we move inside a territory to which we can apply the principle of limits of reason that has been recently claimed – among others – by Noam Chomsky [2].

The second application could do good use of contemporary pluralism regarding free will and try to identify to what extent the confessional boundaries from the past overlap with the present divisions, and some correlations can be traced, for example, between determinism and predestination theology; or compatibilism and theological defence of free will. The old discussion trying to render compatible free will and divine omniscience and omnipotence can be reframed in the current philosophical terms and find clear parallels. This is an exercise not just on anachronistic parallelism and contrast, but an hermeneutic reflection that could trace back in history issues that appear as constant in anthropological study.

b. A More Explicit 'Empirical Stance'

Reading Rahner's analysis on freedom one can feel the strange idealistic and aprioristic style characteristic of the speculative theology, ancient and modern. I have said before that in many cases theologians dealing with freedom were somewhat inspired by their own experience, and not just trying to interpret normative texts from what has been considered as 'divine revelation'. This is true in most cases, but sometimes we can get the impression that theological reflection has abused speculation and has lost sight from the real world and the human and social experience regarding freedom.

Freedom is a traditional topic inside Christian anthropology. In my own experience, this is a treaty that cannot rely only on hermeneutics and tradition, but needs to be updated according to the new contributions provided by auxiliary sciences trying to better understand human nature, otherwise theology would loss reality-contact in its way to know theologically human persons.

The former reflections point to a more 'empirical stance' as the Christian philosopher Bas van Fraassen [12] claimed, in order to render our theological views more updated and significant. Even if such stance has not always helped the ongoing philosophical discussion, in my opinion developments in the empirical study of human behaviour provide excellent inputs to theologians trying to resist the pressure from more reductionist positions. For instance, the studies of Baumeister and colleagues [1] about the role played by conscious mental processes means a blow against those reducing the reach of that faculties, and hence a vindication for free conscious decision making and all its moral implications.

Theology should be aware about these discussions and rely more on empirical evidence provided by scientific approaches, after tests and accurate results when trying to estimate the extent of human freedom and human frailty.

c. Freedom as a Belief

This is possibly the most controversial claim I am doing in my paper. Several ideas formerly exposed point towards this conclusion: the issue of freedom is less related to evidence or cannot be settled by philosophical argument, and it becomes at the end more and more a sort of belief, and hence it could be better understood inside the theoretical framework designed as 'belief studies'. This claim can find support in the unsurpassable pluralism in the past and in present times, and still more in the fact that free will is associated with ideological positions, the most patent, modern liberalism, i.e. with general beliefs and values.

That point should not be received as a surprise: confidence in higher or lower freedom in humans becomes at the end a sort of belief, more or less warranted, but nevertheless a belief, which can assume both versions: the religious and the secular; possibly a kind of transversal dynamic can be described beyond the religious-secular divide line. We can describe which traits can be identified with such belief, or with the contrary, the one that states that we hold a very limited free capacity in our behaviour. Indeed, for many analysts this is one of the last division lines in the contemporary

world: between those who trust freedom and consider worthy to assume the risks associated with it, and those who distrust freedom because of the chaos it entails and threatens.

Placing the issue of 'freedom' inside that very recent research field, and very interdisciplinary, trying to better know how beliefs are acquired, develop, change and get extinct would have only advantages. This is a meta-reflexive move and one that possibly will not solve the conundrums associated with free will, but it would help to better understand our approach to freedom and to analyse it in terms of credences placed besides other beliefs, with whom they interact and form conceptual networks, helping us to transit our world in a very uncertain time.[2]

In a similar vein as happens with free will conceptions, beliefs come in degrees too, or rather, they can be described as an spectrum ranging from lowest to highest intensity or conviction. Again we have to deal in this case more with probabilistic reasoning and Boolean logic than with certainties or apodictic arguments. As with other beliefs, we can analyse in this case too, the factors associated with its acquisition, change and loss; with its increase and decline, and to place it in a broader network connected with other convictions, more or less close, concerning human condition.

The suspicion arising now is that such a manoeuvring could result in a weakening of freedom, reduced to a simple belief. However, the world of beliefs is anything but simple and weak, and this is true when we speak about other beliefs, notably the religious ones. Sometimes people sacrifice everything for their beliefs, and this has happen too in cases where people died to support their belief in freedom. Not a world of certainties, but a world of probabilities and beliefs seem to be the one we are getting acquainted, and perhaps this is good news for free will and less for more deterministic positions.

References

1. Baumeister, R. F., E. J. Masicampo, and K. D. Vohs. Do Conscious Thoughts Cause Behavior? *Annual Review of Psychology* 62, 2011, pp. 331-361.
2. Chomsky, N. *What Kind of Creatures are We?* New York: Columbia University Press 2016.
3. Löwith, K. *Meaning in History: The Theological Implications of the Philosophy of History*, University of Chicago Press, 1949.
4. Milbank, J. *Theology and Social Theory: Beyond Secular Reason*, Oxford, Cambridge, MA: Blackwell, 1990.
5. Milbank, J., C. Pickstock, G. Ward. *Radical Orthodoxy*, London, New York: Routledge, 1999.
6. O'Brien, J. B. *The Paradox of Freedom: An Analysis of Karl Rahner's Theology of Personal Choice*, Georgtown University Press, 1997.
7. Oviedo, L. El nuevo estudio científico de las creencias y la teología de la fe, *Telmus* 9-10, 2016-2017, pp. 51-65.
8. Rahner, K. Freedom, II Theological, In K. Rahner (ed.), *Encyclopedia of Theology: A Concise Sacramentum Mundi*, New York: Burns and Oathes, 1975.
9. Rahner, K. Theology of Freedom, In *Theological Investigations* (New York: Crossroad, 1982) 6.178-96, at 181, the German original, „Theologie der Freiheit," is found in *Schriften zur Theologie* (Einsiedeln: Benziger, 1965) 6.215-37, at 218.
10. Taylor, Ch. *A Secular Age*, Cambridge, MA: Harvard University Press, 2007.
11. Taylor, Ch. *Sources of the Self: The Making of Modern Identity*, Cambridge: University Press, 1989.
12. van Fraassen, B. *The Empirical Stance*, New Haven: Yale University Press, 2004.
13. Visala, A. Free Will, Moral Responsibility and the Sciences: A Brief Overview, *ESSSAT News and Reviews* 26-3, 2016, pp. 5-19.

Notes

1. John Milbank and other theologians following him have been very vocal in denouncing modernity failures and the perversion of liberal models of free will.
2. For an account and review on recent 'belief studies' see: [7].

Studia Humana
Volume 8:1 (2019), pp. 75—83
DOI: 10.2478/sh-2019-0005

Are Design Beliefs Safe?

Hans Van Eyghen

VU Amsterdam,
The Netherlands

e-mail: hansvaneyghen@gmail.com

Abstract:
Recently, Del Ratzsch proposed a new version of the design argument. He argues that belief in a designer is often formed non-inferentially, much like perceptual beliefs, rather than formed by explicit reasoning. Ratzsch traces his argument back to Thomas Reid (1710-1796) who argues that beliefs formed in this way are also justified. In this paper, I investigate whether design beliefs that are formed in this way can be regarded as knowledge. For this purpose, I look closer to recent scientific study of how design beliefs are formed. I argue that the science strongly suggest that people easily form false beliefs. As a result, design beliefs can only constitute knowledge if subjects have additional reasons or evidence for design.
Keywords: design argument, cognitive science of religion, safety condition for knowledge.

1. Introduction

Recently, Del Ratzsch proposed a new version of the design argument. He argues that belief in a designer is often formed non-inferentially, much like perceptual beliefs, rather than formed by explicit reasoning. Ratzsch traces his argument back to Thomas Reid (1710-1796) who argues that beliefs formed in this way are justified. In this paper, I investigate whether design beliefs that are formed in this way can be regarded as knowledge. For this purpose, I look closer to recent scientific study of how design beliefs are formed. I argue that the science strongly suggest that people easily form false beliefs. As a result, design beliefs can only constitute knowledge if subjects have additional reasons or evidence for design.

2. Perceiving Design

Philosophy of religion (both contemporary and historical) knows a wide variety of design arguments.[1] They share a common core in which complexity is argued to point to the existence of a supernatural creator. A classic example is William Paley's argument based on apparent design in nature [17].

Contemporary examples argue that a designer or creator best explains the fine-tuning of physical constants needed for human life [11]. Del Ratzsch proposes a rather different design argument.[2] He proposes that people come to hold design beliefs not by means of an inference but by a cognitive process that closer resembles perception. He calls this process 'perceiving design.' In this section, I take a closer look at his argument.

Ratzsch starts off with an observation. He notes that many people have experiences where the belief that x was designed *comes over them* or *happens to them*.[3] The experiences show that acquiring design beliefs is *passive and experiential*. As Ratzsch notes, his line of reasoning is similar to that of Thomas Reid. Reid argued that in some situations certain specific phenomenological content could automatically trigger cognitive states. Although the resulting state follows causally from the phenomenological content, it does not follow inferentially. On these occasions, subjects simply find themselves in a cognitive state. According to Reid, such experiences result from the way the human mind is constituted [26].[4]

Ratzsch not only argues that his description is more in line with how most people form design beliefs, he also suggests that other design arguments piggy-back on perceiving design. He argues that inductive design arguments might depend on a non-inferential process to identify base cases of design. Without being able to identify cases of design, no argument by analogy or induction can get off the ground according to Ratzsch [26]. For example, Paley's design argument where he concludes that nature is designed because nature is analogous to a watch, appears to depend on perceiving design in nature. The argument is only plausible because we are able to intuitively see that the watch is designed and because we intuitively see that nature resembles the watch in its complexity. Ratzsch claims inference to the best explanation arguments (like Holder's fine-tuning argument) might also depend on perceiving design. Judging that design is the best explanation for a phenomenon requires that a subject recognizes some properties of that phenomenon as *design relevant* [26]. For example, the precise alignment of the physical constants in Holder's argument is intuitively recognized as a feature that point towards design.[5] Ratzsch himself does not take a strong stance on whether all design arguments are in the end dependent on perceiving design. It seems as if at least in some cases this is not the case. Holder draws his conclusion after carefully comparing the probabilities of both theism and naturalism given the fine-tuning of physical constants [11]. This goes well beyond a mere intuitive recognition of design or design-like features. Nonetheless, Ratzsch convincingly argues that many people form design beliefs non-inferentially.

According to Reid, the acquisition of design beliefs is similar to the acquisition of beliefs about (other and one's own) minds. He claims that human subjects acquire beliefs about minds only by noting their effects and signs.[6] The connection between signs or effects and minds is simply build into human cognitive architecture. In a similar way, subjects form design beliefs after noting its signs and effects. The signs and effects of design include: contrivance, order, organization, intent, purpose, usefulness, adaptation, aptness/fitness of means to ends, regularity, and beauty [26].

Ratzsch does not discuss whether design beliefs that follow perceptions of design are justified or could constitute knowledge. Some of his references to Reid suggest that he does. When he makes the analogy with acquiring beliefs about minds he quotes Reid as follows: "We are conscious only of the operations of mind in which they are exerted. Indeed, a man comes to *know* his own mental abilities, just as he *knows* another man's, by the effects they produce (…)." ([28] quoted by [26] emphasis added).[7] Reid strongly suggests that perceptions of design can lead to knowledge as well. Ratzsch quotes: "When we consider attentively the works of nature we see *clear indications* of power, wisdom, and goodness." ([28] quoted by [26]). Though Reid is not as firm here, his use of the term 'clear indications' suggests that the works of nature provide strong evidence for knowing that a designer exists.

3. The Epistemic Status of Design Beliefs

Drawing on Ratzsch, Alvin Plantinga discussed the epistemic status of design beliefs formed after perceiving design in more detail.[8] Plantinga argues that design beliefs can constitute knowledge because they are formed in a basic way [22]. Basic beliefs are beliefs that are not accepted on the basis of other beliefs.[9] According to Plantinga, basic beliefs can have warrant (i.e. that quality that makes true belief knowledge) if it is produced by a cognitive process that is properly functioning according to a design plan and is aimed at truth [21].

Plantinga's theory of warrant is not widely accepted. Ratzsch and Reid, however, suggest a more simple way in which design beliefs can be justified and even constitute knowledge. Both suggest that design beliefs are justified because they are similar to how beliefs about minds are formed. In both cases, a subject picks up signs and intuitively forms a belief. In the case of minds, the signs will mostly be external behavior like facial expressions. In the case of design, the signs are apparent order or complexity. We noted that Reid claims that beliefs about minds can constitute knowledge. Since he claims that design belief is similar, he thereby strongly suggest that they con constitute knowledge as well. This line of reasoning is in line with Reid's defense of common sense.[10] Reid defends the validity of common sense judgments. He does not claim that all common sense beliefs are justified but argues that certain common sense principles, which possess the consent of many people, should be considered good ways of forming beliefs. The fact that these principles enjoy widespread consent reveals that they are part of the general human cognitive make-up. Reid argues that these general common sense principles provide good evidence for the beliefs they produce. Reid suggests that the way humans form beliefs about minds and about design are examples of general common sense principles. He thereby strongly suggests that the beliefs they produce are justified.

My aim below is not to assess whether design beliefs can be justified but to investigate whether design beliefs (when produced by perceiving design) can constitute knowledge. A first requirement for qualifying as knowledge is that a belief is true. A discussion of whether there is in fact a designer or creator lies beyond the scope of this paper.[11] I will assume for the sake of the argument that design beliefs are true. A second requirement for knowledge is that a belief is justified. We noted above that design beliefs could be justified in a Reidian framework. To qualify as knowledge, most contemporary epistemologists require more than justification. What a true belief requires to qualify as knowledge is subject of much debate. Some recent proposals argue that knowledge poses a modal requirement. One prominent proposal is a safety condition. I discuss this condition in the next section.

4. The Safety Condition for Knowledge

Before we can assess whether design beliefs (if produced by perceiving design) are safe, we need a clear view of the safety condition for knowledge.[12] As Dani Rabinowitz noted the basic idea behind the safety condition for knowledge is: "an agent S knows a true proposition P only if S could not easily have falsely believed P" [25] Being a modal notion, safety is cashed out using possible worlds. A belief P is thus safe if there is no close world surrounding the actual world where P is produced by the same belief forming process at the same time and false.[13] There are thus four factors that remain fixed when assessing safety: the subject, the belief, the time and the belief-forming process. With these factors fixed, safety gauges whether the subject arrives at true beliefs if other features of the world vary.[14]

For our purposes, the subject is a person who forms design beliefs and the time is the moment after perceiving design. The belief under discussion is the belief that there is a creator. The creator can be regarded as a God or an intermediary being and he can be regarded as having created the earth or the universe. The belief-forming process that needs to be held fixed perceives design as it was discussed by Ratzsch. Rabinowitz makes a distinction between fine-grained and course-grained belief-forming

processes in accounts of safety.[15] A process is coarse-grained if described generally or broadly and fine-grained is described in detail. Specifying detail for a belief-forming process raises a problem known as 'the generality problem' [25]. The generality problem was originally raised against reliabilist epistemologies [4] and states that specifying a belief-forming process in greater or lesser detail can affect its reliability. Vision in general can be regarded as a belief-forming process that generates mostly true beliefs and hence is reliable. When the process is limited to perception at great distance, it produces a lot more false beliefs and is unreliable. One defender of the safety account, Timothy Williamson, acknowledges that the safety-condition faces the generality problem [30].[16] The generality problem can be evaded by being clear about the belief-forming mechanism. When the belief-forming mechanism is specified as 'perception at great distance', there is no problem in assessing the safety of beliefs it produces. I will return to this point below

In order to assess safety, we thus need to look at nearby possible worlds where a subject forms the belief that there is a creator after noting order or complexity. For this purpose, I will look closer to recent scientific study of how people come to believe in a creator.

5. Psychology of Perceiving Design

To assess the safety of belief in a creator formed when people perceive design, I will look at recent work in psychology and cognitive science. Perceiving design has been intensely studied by Deborah Kelemen and her team. In this section, I will give an overview of her and related work

Deborah Kelemen argued that children are prone towards 'promiscuous teleology'. She and her team observed that children are prone to give teleological explanations for phenomena where teleology is absent [12], [13]. In a first study, children were shown photographs of living things, non-living things and artifacts. When they were asked what the thing was 'for', whilst explicitly being given the option to answer that they were 'for' nothing, they tended to assign functions to all things, whether they really were 'for something' or not. For example, a lion was reported to be 'for visiting in the zoo' and clouds were 'for raining'. Adults who were subjected to a similar experiment did not show this tendency. In a second study, children were asked if a thing was 'made for' something. Children again showed a stronger tendency to answer that things were made for something than adults. In a third study, children and adults were given a choice between four categories of answers to questions of how something came to be, 'one time accident', 'frequent accident', 'one time intentional' and 'frequent intentional.' Here, children were keener to give intentional answers than adults. Kelemen concluded that children are promiscuously teleological and not selectively teleological like adults [12]. Margaret Evans reported findings, which support the claim that children of both religious and non-religious households display a bias towards intentional accounts of how species originate [8].

The studies mentioned above only attribute promiscuous teleology to young children. Kelemen and her team also found support for the idea that promiscuous teleology does not disappear in adulthood but rather goes dormant and continues to play an implicit role. Especially when adults were asked to answer similar questions like the children in earlier experiments under time pressure they were more error-prone and also showed a preference towards teleological explanations [14]. A study conducted on Romani subjects, with little or no scientific training, showed that they were more likely to endorse purpose-based explanations of non-living entities [2]. Kelemen suggested that science education causes teleological reasoning to recede but not completely vanish [14]. Adults seem to abandon teleological explanations when they learn scientific, material explanations for the phenomena under investigation. The intuitions, however, remain which suggests that for phenomena for which there is no scientific, material explanation adults will still tend to give teleological explanations.

A study on patients with Alzheimer's disease supports the view that the restriction of teleological explanations in adulthood is fragile. The patients were given a choice between mechanistic and

teleological explanations and preferred the latter. The tendency towards teleological explanations thus appears to recede when children acquire beliefs about the causal mechanisms of what was perceived as designed. However, if knowledge of causal mechanisms is affected by Alzheimer's disease, people slip back in systematically and promiscuously preferring teleological explanations [15].

Kelemen's research provides sufficient reasons to think that people frequently err when judging that something is designed. It strongly suggests that people are prone to form false beliefs that things, or beings are designed for some purpose.

All of this raises the question why people are prone to form design beliefs. Kelemen does not address this question. Stewart Guthrie argues that seeing teleology could be a by-product of the detection of intentional agents. Seeing goal orientation is one of the best cues for detecting agents. Since detecting agents is very important for survival (they might be predators), it is evolutionary beneficial to detect too many agents than too few. As a result forming beliefs about agents when none are around is adaptive. Since seeing goal orientation and teleology is a clear indicator of agency it could thus also aid survival to see too much teleology [10]. Having a clear idea about the evolutionary function could help in assessing the safety of design beliefs. If promiscuous teleology served an evolutionary function, it is likely that people will have it in more nearby worlds. There would thus be more nearby worlds in which people will have the same belief-forming mechanism.

6. Is It Safe?

Having a better view of the belief-forming process behind perceiving design, we can now assess whether beliefs produced by perceiving design are safe. The research by Kelemen and her team strongly suggest that design beliefs formed in this way are not safe. It shows that people easily make mistakes when judging that something is designed.

We are not concerned with the safety of all design beliefs. Design arguments, like the argument by Ratzsch, argue for the existence of a creator God. While conclusions of other design arguments (for example that a watch is designed) might be safe, I will argue that this belief is not. I will clarify my argument with the following example:

Alvin walks through a national park. While walking, he sees the beauty of the nature around him. He also sees how many plants show very complex structures and how animals have traits that are well adapted to their environment. After noting all of this, he forms the belief that nature (with all plants and animals included) is designed by God.

Alvin forms the belief that God designed nature. His belief is produced by the process we discussed in section 2. His belief can be true or not. If his belief is false, his belief is evidently not safe. If his belief is true and nature is in fact designed by God, his belief is safe if, and only if, he would not have falsely believed so in most nearby worlds. It appears, however, that he would have done so since there are nearby worlds where nature was not designed by God and where the belief-forming process will still produce the belief that God designed nature. One such nearby world is a world where nature, with all its complexities, arose by strictly naturalistic means. Let us call this world 'world X'. Simon Blackburn describes such a world:

> Science teaches that the cosmos is some fifteen billion years old, almost unimaginably huge, and governed by natural laws that will compel its extinction in some billions more years, although long before that the Earth and the solar system will have been destroyed by the heat death of the sun. Human beings occupy an infinitesimally small fraction of space and time, on the edge of one galaxy among a hundred thousand million or so galaxies. We evolved only because of a number of cosmic accidents, including the extinction of the dinosaurs some sixty-five million years ago. Nature shows us no particular favors: we get

> parasites and diseases and we die, and we are not all that nice to each other. True, we are moderately clever, but our efforts to use our intelligence to make things better for ourselves quite often backfire, and they may do so spectacularly in the near future, from some combination of manmade military, environmental, or genetic disasters [1, p. 29].

Blackburn claims that his description matches the actual world. To assess the safety we assume that Alvin's belief is true and thus that Blackburn's description is false. If we assume that there is a God who designed nature, world X is at least possible. Modal reasoning over God's existence suffers from well-known problems because God is often considered a necessary being. Necessary existence entails that God exists in every possible world, if he exists. This need not be a problem for us. Even if God exists (which we assumed here), it is not necessary true that God designed nature. There is thus a possible world in which God exists and nature arose from strictly naturalistic processes like described by Blackburn. In that world (or those worlds) God could even still have fine-tuned the physical constants. All we need is the possibility of a world where God did not design nature on earth as Alvin believed.

An obvious counterargument is that world X is far removed from the actual world (still assuming that in the actual world nature was designed by God). World X would differ greatly because in it all of nature arose gradually by cosmic accidents and naturalistic evolution while in the actual world nature arose through an act of design by God. Against this counterargument I argue that both worlds are not far apart. Today many theists accept the Darwinian theory of evolution and accept that it can explain order in nature. They usually accept that the theory is naturalistic and can thus explain order in nature without any reference to God.[17] They add to the naturalistic theory that God is the structuring cause of evolution. The only difference between world X and a world where God is the structuring cause of evolution is what drove the evolutionary process. In world X, evolution is driven by coincidences and in the other world by God.

Since only one factor needs to be different between the world Alvin inhabits and world x, there are many nearby possible worlds to Alvin's where his belief is false. It is therefore clear that his belief is not safe.

7. Criticisms

Jeroen de Ridder argues that there is no nearby possible world in which perceiving design will produce false beliefs in cases like Alvin's. He writes:

> Classical theists (…) hold that there wouldn't even have been a universe, let alone evolved intelligent life, were it not for God's creating and sustaining activity. Moreover, proponents of design discourse are also unlikely to grant the more specific assumption that unguided evolution will lead to anything like intelligent beings such as humans. So someone who wants to employ the above line of reasoning to show that there is an undercutting defeater for design beliefs faces the burden of arguing that unguided evolution could produce human beings and, even worse, the burden of arguing that a naturalistic account of the origins of the universe is plausible. Such claims are typically taken for granted by staunch evolutionists and naturalists, but it should be clear that assuming them in the current discussion about the epistemic status of design beliefs begs the question [5].

Like we did, De Ridder's counterargument assumes that there is a designer who designed nature. With this in mind, his claim implies that there is no nearby world with human beings that was not designed by God. We noted above that many theists accept that Darwinian processes can produce nature that is

complex and seems ordered without the need for a designer like God. Being theists, they add that God is the structuring cause of evolution and therefore do not claim that nature arose by naturalistic processes alone in the actual world. Since they acknowledge that evolution can occur naturalistically, they admit that nature with order and intelligent human beings without God as their causes is possible. Claiming that such a world is *possible* does not beg the question against the epistemic status of design beliefs. It would beg the question if one claims that such a world is *actual*. Contrary to De Ridder, it also seems that many proponents of design discourse would acknowledge that unguided evolution *can* produce intelligent beings like humans. Evolutionary biologists argue that the human brain gradually increased in size over millions of years whereby humans became more intelligent. An explanation of how human intelligence arose in terms of gradual evolution of their brains does not refer to God or anything supernatural and is therefore also naturalistic.

Another criticism De Ridder suggests is that we lack sufficient data on which inputs lead to design beliefs and how design beliefs are produced.[18] It could be argued that we lack a clear view of perceiving design and can therefore not assess whether it will produce false beliefs in nearby worlds. The criticism has no force if we restrict the range of possible worlds for assessing safety to worlds in which the subject, her belief-forming process, and the input that leads to the belief in question (i.e. beauty and complexity in nature) remain fixed. We have sufficient data to claim that in most of these worlds, the subject will form design beliefs after seeing beauty and complexity. It seems as if people will often form design beliefs if the perceived beauty and complexity were not caused by supernatural design.

8. Concluding Remarks

In this paper I argued that many design beliefs are not safe. I argued that recent scientific study shows that people easily come to hold false design beliefs. This implies that belief in a creator God formed in a non-inferential way is not safe because there are many nearby worlds in which people will falsely believe that God designed nature.

We noted in section 2 that Ratzsch suggests that many design arguments depend on non-inferentially perceiving design. If this is the case, these arguments might lose some of their force if perceiving design is unsafe. I raised doubts whether all design arguments indeed depend on non-inferentially perceiving design. My conclusion suggests that the more arguments depend on non-inferentially perceiving design, the more they are tainted by the unsafety of design beliefs that are formed in this way. More complex design arguments, like some versions of the fine-tuning argument, will likely not be harmed. More intuitive arguments, like William Paley's analogical argument, will likely be harmed more.

My argument has important ramifications for many common sense design beliefs. It is very likely that many common people form their belief that there is a designer God in a non-inferential way as described by Ratzsch. My argument shows that their beliefs do not constitute knowledge. Subjects in this situation can still bolster their design beliefs by looking for additional evidence or reasons.

References

1. Blackburn, S. *An Unbeautiful Mind*, UNZ.org 2002, http://www.unz.org/Pub/NewRepublic-2002aug05-00029, Accessed on February 14, 2018.
2. Casler, K., Kelemen, D. Developmental Continuity in Teleo-Functional Explanation: Reasoning About Nature Among Romanian Romani Adults, *Journal of Cognition and Development* 9, 2008, pp. 340-362.
3. Comesana, J. Unsafe Knowledge, *Synthese* 146 (3), 2005, pp. 395-404.

4. Conee, E., Feldman, R. The Generality Problem for Reliabilism, *Philosophical Studies* 89, 1998, pp. 1-29.
5. De Ridder, J. Design Discourse and the Cognitive Science of Design, *Philosophia Reformata* 79, 2014 , pp. 37-53.
6. Dembski, W. A. *No Free Lunch: Why Specified Complexity Cannot Be Purchased Without Intelligence*, Lanham: Rowman and little, 2002.
7. Dembski, W. A. *The Design Inference: Eliminating Chance through Small Probabilities*, Cambridge: Cambridge University Press, 1998.
8. Evans, E. M. The Emergence of Beliefs About the Origins of Species in School-Age Children, *Merrill-Palmer Quarterly* 46, 2000, pp. 221-254.
9. Greco, J. How Must Knowledge Be Modally Related to What Is Known? *Philosophical Topics* 26 (1/2), 1999, pp. 373-384.
10. Guthrie, S. E. Religion and Art: A Cognitive and Evolutionary Approach, *Journal for the Study of Religion, Nature & Culture* 9 (3), 2015, pp. 282-311.
11. Rodney Holder, R. D. Fine-Tuning, Multiple Universes and Theism. *Noûs* 36, 2002, pp. 295-312.
12. Kelemen, D. Are Children "Intuitive Theists"?: Reasoning About Purpose and Design in Nature, *Psychological Science* 15, 2004, pp. 295-301.
13. Kelemen, D. The Scope of Teleological Thinking in Preschool Children, *Cognition* 70, 1999, pp. 241-272.
14. Kelemen, D., Rosset, E. The Human Function Compunction: Teleological Explanation in Adults, *Cognition* 111, 2009, pp. 138-143.
15. Lombrozo, T., Kelemen, D., Zaitchik, D. Inferring Design Evidence of a Preference for Teleological Explanations in Patients With Alzheimer's Disease, *Psychological Science* 18, 2007, pp. 999-1006.
16. Nicholas, R., Yaffe, G. Thomas Reid. *The Stanford Encyclopedia of Philosophy*, Winter 2016 Edition.
17. Oppy, G. *Arguing About Gods*, Cambridge: Cambridge University Press, 2006.
18. Paley, W. *Natural Theology or Evidence for the Existence and Attributes of the Deity, Collected from the Appearances of Nature*, Oxford: Oxford University Press, 2006.
19. Philipse, H. *God in the Age of Science: A Critique of Religious Reason*, Oxford: Oxford University Press, 2012.
20. Plantinga, A. Is Belief in God Properly Basic? *Noûs* 15, 1981, pp. 41-51.
21. Plantinga, A. *Warrant and Proper Function*, New York (N.Y.): Oxford University Press, 1993.
22. Plantinga, A. *Where the Conflict Really Lies: Science, Religion, and Naturalism*, Oxford: Oxford University Press, 2011.
23. Pritchard, D. Anti-luck epistemology, *Synthese* 158 (3), 2007.
24. Pritchard, D., Safety-Based Epistemology: Whither Now? *Journal of Philosophical Research* 34, 2009
25. Rabinowitz, D. *The Safety Condition for Knowledge*, Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/safety-c/, Accessed on February 13, 2018.
26. Ratzsch, D. Perceiving Design, In N. A. Manson (ed), *God and Design: the Teleological Argument and Modern Science*, London and New York: Routledge, 2003, pp. 125-145.
27. Ratzsch, D. Koperski, J. *Teleological Arguments for God's Existence*, The Stanford Encyclopedia of Philosophy, Spring 2016 Edition.
28. Reid, T. *Inquiry Into the Human Mind*, 7nd Edition, Edinburgh: Maclachlan & Stewart, 1872.
29. Williamson, T. Knowledge and its limits, Oxford: Oxford University Press, 2002.
30. Williamson, T. Reply to Goldman, In P. Greenough, D. Pritchard (eds), *Williamson on Knowledge*, Oxford: Oxford University Press, 2009, pp. 305-312.

Notes

1. Design arguments are sometimes called ‘teleological arguments.’
2. According to Ratzsch’s line of reasoning, it is problematic to call it an ‘argument.’ Elsewhere, Ratzsch’s line of reasoning is ranked under the teleological arguments [27]. I will refer to it as an argument as well.
3. Ratzsch provides examples of such experiences from the writings of notable sceintists like Charles Darwin and Francis Crick.
4. Ratzsch refers to Thomas Reid’s book *inquiry into the human mind* [28].
5. Ratzsch gives William Dembski’s argument as example [7]. He notes that Dembski writes that identifying patterns and information for eliminating chance needs *insight*. Dembski adds that the logic of discovery at work in this insight is largely a mystery. Ratzsch suggests that the way people perceive design could explain this mystery [26].
6. According to Ratzsch, Reid even claims subjects *know* minds through their effects and signs [26]. I return to this point below
7. Other quotes of Reid also strongly suggest that perceiving design can lead to knowledge of design according to Ratzsch. He writes: “How do I *know* that any man of my acquaintance has understanding? …I see only certain effects, which my judgment leads me to conclude to be marks and tokens of it.” ([28] quoted by [26] emphasis added).
8. Plantinga uses the term ‘design discourse’ instead of ‘perceiving design.’ For reasons of clarity I use Ratzsch’s term.
9. Plantinga famously argues that belief in God could be a proper basic belief. He argues here that criteria for proper basicality are inductive. He claims they should be “argued to and tested by a relevant set of examples” [20].
10. Ryan Nichols and Gideon Yaffe give a good overview of Thomas Reid’s philosophy and hiw view of common sense [16]. My discussion of Reid’s views on common sense is drawn from their overview.
11. As we noted above, defenders of design arguments argue that there is a designer. Others argue that there is no designer (see for example: [17], [19]).
12. Influential versions of the safety conditon for knowledge have been defended by Ernest Sosa [9], Timothy Williamson [29], [24]. My discussion of the safety condition is based on Dani Rabinowitz overview [25]. I do not discuss problems for the safety account, see: [25], [3] and proceed as if the account is true.
13. Greco and Williamson do not explicitly stress that an assessment of safety requires looking at nearby worlds where the belief is produce by the same belief-forming process. Pritchard does when he writes: “S’s belief is safe if and only if in most nearby possible worlds in which S continues to form her belief about the target proposition *in the same way* as in the actual world, and in all very close nearby possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, the belief continues to be true” [23].
14. Rabinowitz uses the term ‘method’ instead of ‘belief-forming process.’ [25].
15. Rabinowitz also makes a distinction between internal and external belief-forming processes. Processes are internal when they are wholly dependent on the subject’s constitution; they are external when they are not. I do not discuss the distinction at length since the process under discussion, perceiving design, is obviously external. The process refers to apparent order or complexity that is perceived as design. This factor is clearly external to the subject.
16. It should be noted that Williamson does not analyse ‘knowledge’ in terms of safety. He does discuss the safety condition at length.
17. A minority rejects this idea and claims that some complexities require reference to a designer. Their position is known as ‘intelligent design’ (see for example [6]).
18. De Ridder writes: “[W]e don’t know exactly which inputs produce design beliefs as outputs or how inputs are converted into outputs” [5].

sciendo

studia humana
QUARTERLY JOURNAL

Studia Humana
Volume 8:1 (2019), pp. 84—92
DOI: 10.2478/sh-2019-0006

Thought Experiments and Novels

Tony Milligan

Department of Theology and Religious Studies,
King's College London, United Kingdom

e-mail: anthony.milligan@kcl.ac.uk

Abstract:
Novels and thought experiments can be pathways to different kinds of knowledge. We may, however, be hard pressed to say exactly what can be learned from novels but not from thought experiments. Headway on this matter can be made by spelling out their respective conditions for epistemic failure. Thought experiments fail in their epistemic role when they neither yield propositional knowledge nor contribute to an argument. They are largely in the business of 'knowing that'. Novels, on the other hand can be an epistemic success by yielding 'knowledge how'. They can help us to improve our competences.
Keywords: thought experiments, knowing how, knowing that, emotion, impoverished narratives

I.

Novels and thought experiments, or at least good instances of both are exercises of the imagination. Often (but not always) they are also fictional, involving counterfactual (imagined) rather than realized, circumstances. Both may escape the charge of fantasy by helping us to attend to real features of the world. My concern here will be to arrive at a rough story about what separates the two. Beyond saying that I am using 'thought experiments' to refer to impoverished narratives (rather than novels), I have no provisional definition to offer detailing what a thought experiment is. Nor is such a definition necessary in order to say something of interest about thought experiments. None of the available definitions look particularly promising. To say that thought experiments are experimental marks no obvious boundary, given that good novels are also in some sense experimental. Marking the distinction by appeal to non-execution is no more successful. If we can make sense of thought experiments as experiments without execution (Sorenson's position) then we can probably do the same for novels.[1]

We may even wonder if there is any point in marking a boundary here at all, especially when there is a senses of 'thought experiments' in which some novels could qualify.[2] To be a thought experiment in the relevant, less restricted, sense just is to be an appeal to some counterfactual circumstance in order to explore and/or help to answer a question. Obvious examples of novels that are experimental in this sense would be most of the works of Dostoevsky and existentialist novels exploring the relation between freedom and anxiety. Other examples of novels that might count as thought experiments in this broad sense are works of science fiction where *what*

if? questions about physics, time travel and freewill are central to the plotline. Some of the works if Stephen Baxter would qualify. A final and more unusual example of novelistic thought experiment is David Lodge's *Thinks*, unusual because thought experiments from Alan Turing, John Searle and Frank Jackson figure directly within the text. The very idea of thought experimentation is central to the narrative.

I do not intend to dispute the claim of such texts to be thought experiments in some sense, but it is just not the sense that interests me here. Instead, I will be concerned with thought experiments in a familiar but more restricted sense which requires that they are puzzle-like and brief. Here, it is their brevity that will principally be my concern. Thought experiments (in the relevant restricted sense) consider some specific scenario and only that scenario, with minor allowances for presentation. As in formal arguments aiming at deductive validity and soundness, the procedure when constructing thought experiments is to allow for some elegance of formulation while avoiding excess of detail. The reader is informed of the salient facts but they are informed of little else. Engaging in thought experimentation of this sort is a characteristic part of certain kinds of science (such as physics of a very theoretical sort) and also of analytic philosophy although both disciplines have skeptics about their methodological value. In the absence of any provisional definition, an example of this type of thought experiment may serve as a clarification of the kind of impoverished narrative that I have in mind:

> **Crusher and flusher:** You enter a room, at one end of the room is a timed and loaded baby crusher, rather like a small textile press, and at the other end of the room is a timed and loaded embryo flusher. You have, and are aware of having, enough time to race over and switch off one device but not both. The crusher threatens a single infant, but the flusher threatens several embryos. What would you do?

This is a question that may have a serious point, but that does not make it a serious question. We would all rescue the baby on pain of moral idiosyncrasy and gross moral failure. And we would do so just so long as the given information is *all* that we have to go on. This thought experiment uses a briefly described scenario (an impoverished narrative) to help us recognize an intuition about what we value most. It is, in Daniel Dennett's familiar but rather awkward terms an 'intuition pump' [5]. It helps bring to the surface or otherwise to articulate a conviction that we may not previously have recognized ourselves to have. Such articulation can be a philosophically significant activity.

More ambitiously, it may be claimed that the above crusher-and-flusher experiment is (or contributes to) a concealed argument, with an implicit conclusion and one or more hidden premises.[3] Someone who claims that each and every individual human embryo has the *same* value as each and every individual post-natal human, on running through the above brief narrative and taking on board its significance, might discover that there is something askew with their viewpoint or at least with an unqualified statement of it. The concealed argument in this particular instance has the structure of a *reductio* (absurd consequences follow from a given set of premises*)* but my case does not require commitment to the view that the argument concealed within the narrative of a TE must always have this structure.

II.

To say that novels are also counterfactual explorations is not to say that they must be in some sense arbitrary, or must lack anything approximating to an internal necessity where all the events are *required* by the narrative. As explorations of counterfactual circumstances, novels may even seem to have the edge over impoverished narratives of the sort involved in thought experimentation. What I have to say will aim to show one respect in which philosophical appeals to thought experiments are methodologically weaker than novels and not just methodologically distinct. My choice of the terminology of 'impoverished' expresses an acceptance that there is indeed a deficit. Martha Nussbaum, who has the works of Proust and Henry James in mind as members of the

relevant contrast class, tries to cash out just what this deficit is by listing the failings of what she calls "schematic philosophers' examples." Their limitations contrast with the success-making features of good novels:

> They almost always lack the particularity, the emotive appeal, the absorbing plottedness, the variety and indeterminacy, of good fiction; they lack too, good fiction's way of making the reader a participant and friend… If the examples do have these features, they will, themselves, be works of literature [13, p. 46].

Although she allows elsewhere that novels *can be* thought experiments I will take it that what Nussbaum means by "Schematic philosophers' examples" will overlap with what I mean by thought experiments in the relevant, puzzling and impoverished narrative sense.[4] One of Nussbaum's points strikes home particularly well. We do not, or do not normally respond emotionally to thought experiments in the way that we do to novels and to literary fiction in general. There is no parallel to the paradox of fiction (the arousal of emotions about non-events) that needs to be dealt with. In the crusher-and-flusher case, life and its destruction are supposedly at issue, just as they are in novels where characters face decisions about abortion. But anyone who reacted with fear, sympathy or pity that was about this particular counterfactual case (and not just causally connected to it) would have to be in a peculiar delusional state. We can offer partial explanations of just why we experience emotions in response to novelistic (and other sorts of) fiction, but the same does not usually seem to apply to thought experiments.

It is conceivable that someone might think this an advantage, i.e. they might hold that thought experiments are in some sense less prone to induce false emotionally-swayed appraisals of what there is. However, with Nussbaum, we may (perhaps more plausibly) be inclined to regard it as a genuine deficit, at least in cases where thought experiments concern moral responsiveness rather than issues of physics. If one holds to a cognitive account of the emotions, whereby they involve beliefs or a belief-like construal of how things stand with humans, this absence of an emotional response may indicate that something somewhere is lacking in thought experiments. If they do not induce emotional responses then it looks like the scenarios that they involve cannot be *realistic enough*, or else that they may be sufficiently realistic but somehow they still manage to induce emotional oversight.

Nussbaum adds a further interesting wrinkle to this picture to the effect that at least some schematic philosophers' examples *may* have enough of the relevant features to count as works of literature (in the relevant restricted sense of 'works of literary fiction') and presumably this would include their having some sort of emotive standing. In support of this claim she footnotes Iris Murdoch's use of examples in *The Sovereignty of Good* [13, p. 46, n. 84]. I will take it that she has Murdoch's case of D and M in mind. (Nothing else in Murdoch's *Sovereignty* fits the bill.)

> ***The Case of D&M***
> A mother, who I shall call M, feels hostility to her daughter-in-law, whom I shall call D. M finds D quite a good-hearted girl, but while not exactly common yet certainly unpolished and lacking in dignity and refinement. D is inclined to be pert and familiar, insufficiently ceremonious, brusque, sometimes positively rude, always tiresomely juvenile. M does not like D's accent or the way D dresses. M feels that her son has married beneath him. Let us assume for the purposes of the example that the mother, who is a very 'correct' person, behaves beautifully to the girl throughout…Thus much for M's first thoughts about D. Time passes, and it could be that M settles down with a hardened sense of grievance and a fixed picture of D…However, the M of the example is an intelligent and well-intentioned person, capable of self criticism, capable of giving careful and just *attention* to an object which confronts her. M tells herself: 'I am old-fashioned and conventional. I may be prejudiced and narrow-minded. I may be snobbish, I am certainly jealous. Let me look again.' Here I assume that M observes D

> or at least reflects deliberatively about D, until gradually her vision of D alters…And as I say, *ex hypothesi*, M's outward behaviour, beautiful from the start, in no way alters [11, pp. 16-17].

This thought experiment can help the reader to articulate or recognize the intuition that we can be active in a morally praiseworthy manner without engaging in publicly-observable behavior. Inner and morally significant events may stand in no need of an outer criterion. D&M also succeeds in the more generous terms that Nussbaum allows. It is an impoverished narrative of a sort that could be enriched and worked into the plotline of a novel. Indeed there is one Murdoch novel (*Bruno's Dream*) in which a reworked version of the scenario does happen to be played out.[5] But even were this not to be the case, D&M already looks like a work of literature in miniature.

Moreover, while sympathy or compassion would be inappropriate responses to crusher-and-flusher, it is not *obviously* the case that we can say the same about D&M. It is intelligible that someone who has considered and weighed-up the example over several years might at least believe themselves to experience a mild degree of sympathy or compassion for M. And while we might question whether their first-person report was accurate, and whether the compassion or sympathy, if present, was *about* M rather than about non-fictional persons in similar predicaments, we could equally well do the same in the case of any emotional response claimed by the reader of a novel [20, pp. 9-10].

Accordingly, when it comes to inducing an emotional response, Nussbaum seems to have good reason to avoid over-generalizing and to allow that some schematic philosophers' examples/thought experiments might make it into the literary fold. D & M looks very different from crusher-and-flusher, less impoverished and more familiar. We can see how it could be embedded within a background of normal life and this is just what we cannot do in the crusher-and-flusher case. It is a scenario that isolates itself off from the normal, richly detailed background of our world, the detail that might legitimate an emotional response. In that sense it is analogous to experimentation under isolated laboratory conditions.

However, while we might at least entertain the idea that suitably constructed thought experiments can induce an emotional response in normal rational agents, and that this response can be in keeping with the rational character of these agents, it would be odd if thought experiments were to expand the emotional repertoire of such agents in any direct manner. And this is something that novels seem to be capable of doing [14, pp. 236-237]. Someone who lacked compassion could begin to grow through their encounter with literary characters, through their becoming familiar *for the first time* with the kind of joined-up narrative on which so much of the experience of compassion depends. It is a familiar point made by both Nussbaum and Murdoch that when encountering characters in novels we take the time to attend that we often do not take with the individuals that we encounter in everyday life. This may be, in part, because of the differences between novels and everyday life: it is easier for us to attend when there is less at stake and when the overall experience is likely to be enjoyable. But this is often how we learn, in the easier context first.

Be that as it may, whatever the limits of the compassion between attending to a character in a novel and attending to a non-fictional other, novels still seem to make possible the cultivation of a new pattern of emotional response through the *patient* disclosing of details that we would ordinarily, impatiently or inattentively overlook. Allowing for the possibility of some rare exceptions, even a thought experiment such as D & M that provides scope for genuine emotional response, is likely to do so by drawing upon an existing pattern of emotional response and an existing repertoire of emotions, rather than by expanding our repertoire or by otherwise altering our pattern of response.

III.

This is a difference that I take to hold in most cases, *on the whole*, or *generally speaking*. It is not

itself the loose boundary marker that I want to situate but it may be symptomatic of the fact that it *is* appropriate to place some boundary marker between novels and thought experiments. The marker itself will be set down as follows. Novels and thought experiments, or rather good instances of either, can play an educative i.e. knowledge-generating role. But in the case of novels what we learn can be *competences or skills* for encountering the other, and this is something rather different from the propositional knowledge that both thought experiments and novels can help us to acquire.

The difference may perhaps be better appreciated if we reflect upon the respective conditions for epistemic failure of the two kinds of narrative. In the case of thought experiments the conditions for epistemic failure are, up to a point, clear cut. A thought experiment fails when it neither helps to articulate some intuition nor functions as a concealed argument (or as a contribution to such an argument). And to say this much is to make it clear that thought experiments are primarily concerned with *knowing that* something is the case. Moreover, this *knowledge that* usually concerns something general. It is *not* just relevant to the peculiar circumstances that the experiment happens to specify.[6] We are not ultimately concerned to *know that* in the crusher-and-flusher scenario we would do or ought to do one thing rather than another. We are concerned to *know that* on pain of moral idiosyncrasy we value or ought to value individual infants more than collections of individual embryos and that we should act accordingly.

My point here is that thought experiments in moral contexts, but not only in moral contexts, can allow us to make rule-like generalizations. And it is this generality that helps to explain why those who stress the importance of the particular (again Martha Nussbaum, Cora Diamond and Iris Murdoch) together with those who reject guidance by moral principles (particularists such as Jonathan Dancy) are also critics of appeals to schematic thought experiments in moral contexts [3]. As a first approximation we can say that when a thought experiment does not function as, or contribute to, a concealed argument *and* fails to yield (i.e. to promote our identification and acceptance of) some appropriate generalizing proposition it is an epistemic failure. And here we need not require that the thought experiment yields what its author claims that it yields. An experiment that fails to show what its author claims is not thereby automatically a failure but if it fails to contribute to *any* argument or to yield *any* appropriate generalizing proposition then it *is* an epistemic failure.

As a slightly modified version of the above claim we might allow that thought experiments sometimes help us to learn new concepts or to refine existing concepts rather than, or as well as, helping us to acquire *knowledge that*. According to Kuhn some thought experiments do not yield propositional knowledge but help to generate a paradigm shift involving some form of conceptual change or refinement. [7] This may be a rare matter, but I see no reason to deny that something of this sort may from time to time occur. If we accept this, and accept that conceptual acquisition or refinement can count as knowledge acquisition, then our account of the conditions for epistemic failure will take the form of a conjunction. Where a thought experiment is not a concealed argument or a contribution to such an argument *and* where it also fails to pump out or give reason for some appropriate generalizing proposition *and* has no disposition or tendency to improve our conceptual repertoire *then* it is an epistemic failure. And while it might succeed in some other way, for example as a source of amusement or as a good way to loose some time on a train, it is not a success *as a thought experiment*.

If we assume that novels play an *educative* role they too must have conditions for epistemic success and failure. But these conditions are not simply a matter of failing to yield or contribute to *knowledge that*. Indeed, a good novel by an analytic philosopher (if we can stretch our imaginations far enough to allow for such a possibility) might well be one in which the author looses sight of her usual concern to sharpen up intuitions and to generate sound or valid arguments. Novelistic success for such a remarkable being might come at the price of failing to deliver just what a thought experiment must usually deliver if it is to be a successful thought experiment.

Nevertheless, as they are works of the imagination, novels which fail comprehensively from an epistemic point of view, which fail in any important sense to help the reader attend to real features of the world, are works of sheer fantasy. (Here I use ‘fantasy’ in a way that contrasts with

imagination and is different from its use in the shelving classification 'fantasy literature'). Good novels open up possibilities of knowledge, but often it is knowledge of a different sort from that promoted by consideration of a thought experiment. Sometimes, and perhaps often, these two kinds of narrative are not rival pathways to the same thing. It would, for example, be odd under many circumstances to turn to a novel to gain certain kinds of non-trivial *knowledge that*. We might of course, turn to a novel to get trivial knowledge or knowledge concerning the novel itself. Someone might have to *get up to speed* on a particular text for a lecture or exam, or for typesetting or statistical purposes they might want to know the exact number of times the letter 'q' is used in a text. But setting aside such matters, novels are not systematically in the business of making available this kind of *knowing that*. If we want to *know that* utilitarianism involves claims x and y we might turn to Dickens' *Hard Times*, but we would not usually do so. Nor would we usually recommend this procedure to others. Perhaps more importantly, we could not defensibly say that it was the kind of novel from which we can learn nothing of philosophical importance if its account of utilitarianism happens to be skewed or otherwise uncharitable (which it is).

This is not to say that we can never gain *knowledge that* from novels but rather to say that acquisition of this kind of knowledge is often something of a bonus. If we want to *know that* such and such an event happened at the battle of Austerlitz we would not normally turn to Tolstoy, but we might turn to the death of Petya Rostov in *War and Peace* or to *Ivan Ilych* if we want to *know that* encountering death can involve experiences of some particular kind. And here *knowing that* and conceptual acquisition or refinement may go hand in hand. We may learn something about what grief involves by reading the novels in question. And it may even be that conceptual refinement is a regular part of reading novels of the best sort in a way that is not the case for thought experimentation. Nevertheless, we can appropriately regard a novel as an epistemic success without making any appeals to *knowledge that* or to conceptual acquisition and refinement. Even if we come away with no new information, no new propositional attitudes, or with no deeper understanding of some concept, we may still have *learned* something.

What I want to suggest is that the other sort of knowledge made available by novels is the same type of knowledge that *may* arguably be gleaned from the most interesting mystical texts. They may contain all sorts of literal falsehoods, ambiguities and occasional nonsense, but some of those who study such texts do seem to be unusually competent humans and do seem to have learned something. Although, here we might wonder if their competences are in part the result of reading the texts or if they diligently read the texts because of competences that they already possess. (I make this claim only as a way of putting matters that will clarify matters for some readers.)

A similar claim about the acquisition of competences, perhaps a less controversial one, but still controversial up to a point, may be made in relation to texts by Derrida.[8] Once we have a grasp of the big themes that are repeatedly worked upon and may be set out more or less lucidly in propositional form, we can still learn something else from such texts but we might be hard pressed to say exactly *what* we learn. I want to suggest that in such cases what we can gain, or what the attentive diligent reader can gain, is *knowledge how* by contrast with familiar forms of *knowledge that*.

Reading a mystical tract, or Derrida (and these two activities may sometimes be one and the same) *can* help to teach us how to encounter uncertainty and indeterminacy with fresh eyes. That is to say, it can improve the quality of our practical reason with the latter understood in a sufficiently rich sense and not reduced to an ability to terminate arguments with actions. We can gain knowledge of *how* to cope with, how to realistically encounter tracts of experience that do not offer themselves up with clear signposts or assembly instructions. Similarly, and here I make a familiar point that is associated at least with Nussbaum and Murdoch, both the authoring and the reading of novels involves a succession of exercises of attention. The reading of good novels can educate us in *how to attend* to the particularity of what is other, *how* to do so patiently and justly. If we are unwilling to engage with characters who are in some respect unattractive we will not get the enjoyment that we otherwise would from the reading experience. A good novel is one that may involve moral ambiguity so that we cannot just privately *boo* the villains and *cheer* the heroes. It

may also require that the characters have genuine particularity and are more than recognizable stereotypes for some or other character trait, virtue or flaw.[9]

Here, as opposed to the case of Derrida's texts where an instructive mysticism is in play, we are clearly in territory where the gain in knowledge is of a morally relevant sort. It is *knowledge how* of a sort that may figure in moral education. Attending patiently to flawed individuals (and we are all flawed individuals) is *not* something that we ordinarily do well, or that we *know how* to do by default. To *know how* to do this is an epistemic accomplishment that is not an ordinary part of our socialization. Moreover, learning to respond to particularity is itself, in one sense, also a general competence, something that can be carried from situation to situation. And in this respect we may be said to learn something of general applicability in the case of diligently reading novels, just as we learn something general (often of a different sort) from a thought experiment. In the one case we may acquire a general principle and in the other a generally applicable skill.

To state matters in a more concrete manner consider the following situation. A teenager is given a copy of Dickens' *Little Dorrit* and that this is their first big book. They might come to admire the supply of patience that the little seamstress Amy Dorrit appears to have. They might come to learn the general lesson *that* patience is or can be a very admirable thing and that it is required if we are to do justice by the other. However, this is not at all the same as actually becoming more patient as a result of reading the book. Yet the latter is also a possibility. *Little Dorrit* is, after all, a substantial volume and a good deal of the detail is not strictly helpful to the plot. To get through it the teenager may have to effortfully *stick at it* and not give up. And it is in the application of this effort that they may learn to become more patient and also learn to self-trust with regard to seeing things through to the end. As well as this, there is a sense in which such a reader may also enrich their grasp of what patience involves. And these epistemic gains are different from learning *that* patience is or can be admirable. Indeed, the gains might equally well might be made by reading a large and challenging novel in which patience does not figure directly as a theme. But to say this is not to claim that a long novel cannot fail to cultivate patience. Whether it does so or not will depend, in part, upon its being sufficiently engaging to give the reader some reason to keep going. (Generations of readers have happened to find *Little Dorrit* engaging in the relevant manner. Proust's *Recherche* would be an extreme example of patience cultivated and rewarded.) And so, there may be an intimate connection between novelistic content and the potential of any novel to help the reader acquire knowledge how.

IV.

The above goes some way towards giving an account of what it is that good novels do and that thought experiments do not do. Literature helps to improve the quality of our practical reason and this gives some justification (if justification were ever needed) for the claim that philosophers, or at least moral philosophers, ought to read and engage philosophically with novels as a regular part of their philosophical practice. However, I have set up this contrast between novels and thought experiments by appeal to a *knowing how/knowing that* distinction which may seem vulnerable to familiar suspicions concerning overly-clean and clear-cut distinctions. This suspicion is at work, for example, when Iris Murdoch writes against a demand for precision where a tolerance for ambiguity may be less misleading [10]. Derrida writes in a similar, but more systematic vein, about the danger of shoring-up implausibly hygienic binary contrasts of a familiar and problematic sort, the sort that privilege one side of the contrast but then covertly rely upon the other side [6, pp. 41-42].

Here, I think that we can make a move that Hilary Putnam favors and highlight a difference between what we may call 'dichotomies' and what we may call 'distinctions'.[10] Dichotomies purport to have complete generality. They may be rolled out across an entire field of enquiry. Distinctions claim no such dominion. They allow for a role to be played by classifying cases as instances of *this* or *that* but they also allow that there will be blurring in some or even many cases. And the contrast between distinctions and dichotomies may itself be of this sort. I will take it that what the Derridean critique legitimately targets is rigid dichotomies, and that rejection of these does

not entail rejection of the distinctions that apply *for the most part* or only *up to a point*. In line with this, my appeal to a *knowing how*/*knowing that* distinction is not a suitable target for a Derridean critique.

Moreover, insofar as there is a problem of privileging one side of the contrast, it is the danger of privileging *knowing that*.[11] Instead, I am trying to affirm the distinctive moral importance (which is not to say greater importance) of the kind of *knowing how* that novels can help us to cultivate. Indeed, when we think about morality, while it may be the case (as Murdoch's D&M example suggests) that actions are not all that matters it is nonetheless still the case that actions do matter a great deal. Practical wisdom and competence to perform, both of which involve *knowledge how*, is not, from a moral point of view second rate.

With the distinction still available, and with *knowledge how* accorded appropriate standing, we can cash out one important feature of the contrast between thought experiments and novels. Thought experiments are (generally) geared to the production of propositional knowledge while novels, may yield *knowledge that* but they are also effective in the production of *knowledge how*, where the latter is to be understood as knowledge of a practical sort that happens to fit novels particularly well to the task of moral education.

A qualification here concerns one area of the overlap, where both kinds of narrative assist us in the refinement of our concepts. We might wonder about just what it is to gain *knowledge how* and if it is anything other than a form of conceptual knowledge. If it reduces to the latter then insofar as thought experiments are regularly capable of conceptual refinement there will be a kind of *knowledge how* that they will also be regularly capable of providing. But such a reduction does not look attractive. The following may be said against it: if I learn how to encounter others more realistically through the reading of novels, or in any other way, then I am, in a sense refining my grasp of what it is to be human and perhaps my grasp of various other (thick) concepts as well. I may deepen my grasp of what is involved in being *just*, *generous* or *humble*. But, on the other hand, if I learn how to ride a bike it is not so obvious that my concepts need to have altered. (Unless we have a very reductionist and dispositional account of what is involved in the mastery of a concept.) The former example suggests that acquisition of *knowledge how* may (and perhaps may often) involve conceptual acquisition or refinement. The latter case suggests that this is not always the case and that acquisition of *knowing how* cannot reduce to conceptual acquisition or refinement.

The upshot is that even when we allow that both novels and thought experiments may help us to make conceptual progress there still remains something significantly different about the range of epistemic roles that these differing kinds of narrative characteristically play. And this something different involves novels having a wider range of conditions for epistemic success or failure.

References

1. Bishop, M. Why Thought Experiments are Not Arguments *Philosophy of Science* 66, 1999).
2. Brown, J. R. *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*, London: Routledge, 1991.
3. Dancy, J. The Role of Imaginary Cases in Ethics, *Pacific Philosophical Quarterly* 66, 1985.
4. Davies, D. Thought Experiments and Fictional Narratives, *Croatian Journal of Philosophy* 19, 2007.
5. Dennett, D. Intuition Pumps, In J. Brockman, *The Third Culture: Beyond the Scientific Revolution*, Simon & Schuster: New York, 1995.
6. Derrida, J. *Positions*, Chicago: Chicago University Press, 1982.
7. Kuhn, T. A Function for Thought Experiments, In T. S. Kuhn, *The Essential Tension*, Chicago: University of Chicago Press, 1977.
8. Moore, A. W. Arguing with Derrida, *Ratio* 13 (4), 2000.
9. Murdoch, I. *Sartre, Romantic Rationalist*, London: Vintage, 1999.
10. Murdoch, I. Against Dryness, In I. Murdoch *Existentialists and Mystics: Writings on Philosophy and Literature*, P. Conradi (ed.), London: Penguin, 1999

11. Murdoch, I. *The Sovereignty of Good*, London: Routledge, 2001.
12. Norton, J. Are Thought Experiments Just What You Always Thought? *Canadian Journal of Philosophy* 26, 1996.
13. Nussbaum, M. *Love's Knowledge*, Oxford: Oxford University Press, 1992.
14. Nussbaum, M. *Upheavals of Thought*, Cambridge: Cambridge University Press, 2001.
15. Putnam, H. *The Collapse of the Fact/Value Dichotomy*, Cambridge Mass.: Harvard University Press, 2002.
16. Regan, T. *The Case for Animal Rights*, Berkeley: University of California Press, 2004.
17. Sorenson, R. *Thought Experiments*, Oxford: Oxford University Press, 1992.
18. Stanley, J., and T. Williamson. Knowing How, *Journal of Philosophy* 98 (8), 2001.
19. Swirski, P. *Of Literature and Knowledge: Explorations in Narrative Thought Experiments, Evolution and Game Theory*, London: Routledge, 2006.
20. Walton, K. Fearing Fictions, *Journal of Philosophy* 75 (1), 1978.

Notes

1. According to Sorenson [17, p.205], a thought experiment is 'an experiment that purports to achieve its aims without the benefit of execution', but this would leave the genuinely experimental status of thought experiments open to question.
2. For a quite different treatment of extended fictional narratives as thought experiments (in a less constrained sense) see [4] and [19].
3. For the view that thought experiments are really arguments see [1] and [12].
4. Nussbaum's essay on 'Transcending Humanity' in [13] uses 'thought experiment' in a wider sense that is inclusive of novels and related works such as Homer's *Odyssey*.
5. In Murdoch's 1969 novel *Bruno's Dream* a father-in-law is confined to his deathbed and unable to do anything but reflect upon matters. He has to face his failure to welcome his daughter-in-law while there was still time, before her tragic death and his son's subsequent estrangement.
6. To allow that the knowledge in question does not simply concern the circumstances specified in a particular experiment is consistent with allowing various restrictions of the sort set up by Tom Regan to the effect that some thought experiments tell us what to do in exceptional cases and not in normal cases [16, pp. xxvii-xxx].
7. Kuhn [7] and for a contrasting classification of thought experiments as primarily concerned with theory refutation and support, see [2].
8. Moore [8, p.367 ff.], approaches Derrida from the standpoint that of 'knowing how' rather than 'knowing that' albeit his reason for doing so concerns the ineffable. I have no objection to this but an appeal to ineffability plays no direct part in my approach to the novel.
9. For the limited value of using character types in novels see Iris Murdoch's account of Sartre's *Roads to Freedom* trilogy in [9, p.56 ff].
10. See [15, pp. 9-11] for the contrast between 'distinctions' and 'dichotomies'.
11. This suspicion about *knowing how* is particularly evident in [18].

sciendo

studia humana
QUARTERLY JOURNAL

Studia Humana
Volume 8:1 (2019), pp. 93—98
DOI: 10.2478/sh-2019-0007

Biology and Gettier's Paradox

Gonzalo Munévar

Professor Emeritus, Lawrence
Technological University, Michigan, USA

e-mail: gmunevar@ltu.edu

Abstract:
Gettier's Paradox is considered a most critical problem for the presumably obvious philosophical view that knowledge is justified true belief. Such a view of knowledge, however, exposes the poverty of analytic philosophy. It wrongly assumes, for example, that knowledge must be conscious and explicit, and, to make matters worse, linguistic, as illustrated in Donald Davidson's writings. To show why this philosophical view is wrong I will point to arguments by Ruth Barcan Marcus and, principally, Paul Churchland, as well as to work by the neuroscientist Paul Reber on intuitive knowledge. We will see, then, that much of our knowledge is neither explicit nor conscious, let alone linguistic.
I will suggest that an approach that pays attention to biology is more likely to succeed in developing a proper account of our cognitive abilities. Thus, Gettier's paradox becomes a mere curiosity.
Keywords: Gettier's paradox, justified true belief, non-linguistic knowledge, intrinsic learning, neural nets.

1. Introduction

A biological approach to knowledge provides philosophers with a promising alternative to analytic epistemology. For example, philosophical analysis recognizes as intuitive the notion that knowledge is justified true belief. That intuition, unfortunately, runs into trouble because of Gettier's paradox, but analytic philosophers, far from being professionally embarrassed, revel in the opportunity to either solve the paradox or make it even more perplexing. But to me, the main problem with the apparently intuitive notion that knowledge is justified true belief is that it assumes that knowledge is propositional and thus linguistic. This assumption is much at odds with evolutionary biology and recent advances in neuroscience. Many philosophers defend the autonomy of philosophy against such scientific interlopers, but it seems to me that the case for philosophical autonomy, at least where it concerns the issue of knowledge, is weak and implausible. If my arguments are accepted, the most important paradox of contemporary analytic philosophy should become little more than a scholastic curiosity. Indeed, philosophical analysis, at least in its linguistic mode, will become little more than an occasionally useful tool in an epistemology more in consonance with our scientific times.

2. The "Most Significant Problem in Epistemology"

The great majority of analytic philosophers consider Gettier's paradox to be the most significant problem in epistemology. The paradox goes as follows. Let us say that Mary knows that Paul is in the study. According to philosophical analysis, this claim presumably means that
(1) Mary believes that Paul is in the study.
(2) Mary's belief is true.
(3) Mary is justified in holding her belief.

What counts as justification may be a matter of debate – perhaps all it takes is for Mary to see Paul in the study – but as long as we agree that she is indeed justified in her belief, and that her belief is true, we should conclude that Mary knows that Paul is in the study.

Imagine the following situation, however. There is a perfect replica of Paul "sitting" at the desk in the study. Looking through the window, Mary sees the replica and forms the belief that Paul is in the study. Presumably this is a justified belief. But, and this is Gettier's trick, imagine also that Paul is in the study. Not at the desk, but hiding behind the couch. Still, the sentence "Paul is in the study" is true. Thus, Mary believes that Paul is in the study, her belief is justified, and her belief is true. But surely, we want to resist concluding that she *knows* that Paul is in the study.

Analytic philosophers have proposed a variety of ways to solve this paradox, most of which have caused much debate. A popular move, for example, is to demand that in addition to justified true belief certain other conditions be met before we consider that a particular claim constitutes knowledge ("JTB Plus"). But discussing such moves is not my concern in this paper.

My suggestion is that analysis does not settle the matter. Indeed, I will go further than that: Analysis is the wrong approach to determine the nature of knowledge.

Philosophers from A-Z have occasionally found the JTB account unintuitive. Just to mention the P's, remember that Plato in his *Republic* thought that knowledge and belief (opinion) were so different in kind that no qualification could possibly make a belief count as knowledge. Thus, for him, having a belief could not be a requirement for knowledge. And Popper argued that scientific knowledge could not involve justification (in the way philosophers think of justification). Science works by trial and error: Scientists propose hypotheses and try to falsify them. Persistent failure to falsify a hypothesis does not justify it; at best, it inclines scientists to accept it tentatively (the next test may finally refute it). Important philosophers have thus thought of knowledge without belief or justification.

Of course, they could be wrong while analytic philosophers are right in wringing their hands about Gettier's apparently unsolvable paradox.

Plato, Popper, and the analytic philosophers nonetheless seem to agree on a crucial connection between knowledge and rationality. For analytic philosophers, rationality tends to be defined in terms of consistency, implication, logical truth, etc., of sentences or propositions. A rational agent approves of consistent sentences, for example, and strongly disapproves of contradictions. Knowledge is linguistic, and so is belief. Creatures without language, Donald Davidson argues, cannot have beliefs (and thus cannot have knowledge). Moreover, he asks, "Can a creature have a belief if it does not have the concept of a belief?" The answer is no, apparently because creatures without language can have no concepts. As Davidson explains further, "Someone cannot have a belief unless he understands the possibility of being mistaken and this requires grasping the contrast between truth and error – true belief and false belief. But this contrast… can emerge only in the context of interpretation [of a language]" [5, pp. 22-23]. This is not surprising, since truth and falsity are properties of sentences (or propositions).

This implies that dogs, chimps and young children have no beliefs, and thus no knowledge, since they are not language users. But denying them beliefs seems absurd, as Ruth Barcan Marcus argues [10, pp. 233-256]. Consider a case, she suggests, in which Jean and his dog Fido are lost in the desert. At one point they see a mirage of an oasis and they crawl eagerly towards it. Their behavior makes it reasonable to say that they both have the (mistaken) belief that there is water a few meters in front of them. But according to the likes of Davidson, Jean believes mistakenly that there is an oasis a few meters in front of him. Fido has no beliefs at all [10, p. 234]. To make matters worse, for Davidson and others, belief is a conscious relation between a subject and a sentence. This would rule out all unconscious, or subconscious, beliefs. This is, again, unreasonable. As Barcan Marcus points out, being asked why we act as we do may make us realize, for the first time that we have certain beliefs, indeed we might have had them for a while even though we were not aware of them. We do not always "entertain propositions or sentences we hold true while acting." For example, "I often walk a route to my office that is not the shortest and am asked why. It requires some thought. It isn't out of habit, I decide. I finally realize that I believe it to be the most scenic route" [10, p. 239].

Split-brain experiments, and a great number of other experiments in neuroscience clearly indicate that we have unconscious beliefs. Moreover, the brain mechanisms involved are not found only in humans. In actuality, several animals also have the brain structures apparently involved in some of our conscious experiences.

The linguistic "imperative" when it comes to belief seems quite feeble now. But if belief can be non-linguistic, so can knowledge. If knowledge is made of the "right kind" of beliefs (i.e. justified), but those beliefs turn out not to be linguistic, knowledge will also not be linguistic.

Perhaps we could insist that only linguistic beliefs can count as knowledge. But consider that without a decent grasp of their environment, including their social environment, many animals would be unable to function and survive. Why should we say that such a grasp does not amount to knowledge? Indeed, knowledge can clearly be adaptive for many creatures. A chimp, for example, will track down ants to their colony. He will then break a branch off a bush, clear it of leaves, smear it with saliva, and poke it into the entrance to the anthill. From time to time he will take this convenient tool out and eat the ants that have got stuck to it.

But if animal knowledge is not linguistic, so is much of human knowledge. A gifted football (soccer) player (in Spanish: "de los que saben" – one of those who know) will instantly grasp the lay of the field and will give the ball the right touch so it will curve over and around opponents and land at the feet of a sprinting teammate with a chance to score. If the gifted player stopped to think consciously about he was to do, his play would fall apart. Conscious verbalization, since it takes even longer, would likely interfere with his knowledgeable behavior even more.

For some this is a case of "knowing how," not of the relevant "knowing that." But let us say that I am very good at reading people, at least certain people, and tell whether they are lying or not. Liz sits in front of me and gives me an excuse. Just from my unconscious (or subconscious) reading of her I can tell whether she is lying or not. But if I try to consciously verbalize the workings of my brain in picking up her clues, I lose my chance of being able to tell *that* she is lying (or *that* she is not).

The neuroscientist Paul Reber offers a very telling example:

> A fireman in Cleveland cleared his team from a fire scene because he "sensed" that something was odd about the situation. Indeed, the floor was about to collapse because of a raging fire below. The lieutenant fireman who saved his men was not aware of the danger in the usual sense, but rather he was observant enough and skilled enough to know that something was not right. He acted on that indication before consciously realizing what wasn't right or what danger was present. At first he thought it was ESP. Only much later did he begin to understand the clues he had sensed.

This example of successful intuitive knowledge, Reber tells us, "can be credited to implicit processing of the environmental cues, leading to escape from an imminent catastrophe… our brains possess an array of mechanisms for automatically extracting information from the environment without our awareness." It is his conjecture, thus, that "implicit memory is critical in producing trustworthy intuition" [12, pp. 474-475]. Reber tells us that deliberate processing can actively block the use of intuitive knowledge, as we can see in the football player example given above. The mechanisms of implicit learning may also interfere with conscious reasoning, and "the systems often appear to compete such that only one system can influence behavior." Nevertheless, sometimes they do cooperate, e.g. in as fundamental a cognitive activity as categorization. Indeed, as Reber informs us, extensive neuroscientific investigations have even revealed the key brain regions involved: medial temporal lobe activity is associated with explicit memory for prior examples; posterior caudate activity correlates key brain systems associated with implicit learning; and dorsolateral prefrontal cortex activity is associated with resolving competition between implicit and explicit processing [12, p. 479].

The notions of knowledge entertained by analytic philosophers do not seem to do justice to our cognitive abilities, let alone those of animals. Now, if I may be allowed a personal anecdote, after years of doing research and teaching cognitive neuroscience and cognitive psychology, as well as other related courses, I decided to look at the many textbooks I had used, or considered using, or had reviewed for publishers, just to see how important the notion of "belief" was to the science of human cognition. The first step in such a search is to look for "belief" in the subject index. I was not able to find any appearances of that word in any of those books. Perhaps I missed one or two, but I doubt it. In science, the notion of belief, let alone justified true belief, is hardly ever used to investigate the nature of knowledge.

In the *Theaetetus*, Plato tells us that to make a true judgment about something we must already be able to distinguish it from other things (209a-b). If someone can always, or nearly always make the right distinctions, why should that ability not suffice for knowledge? (Why must we also have an "account," as Plato put it, or a "justification," as analytic philosophers put it?)

Knowledge can be demonstrated – and I think we can agree with the analytic philosophers on this – when the agent almost unerringly makes the appropriate conceptual distinctions. It is a fiction, however, to hold that language is necessary for having concepts. Vectorial transformation of information in the brain, for example, explains how concepts are located in non-linguistic vectorial spaces. What is not located in a vectorial space is taken to be different from the concept in question. Paul Churchland points out that our ability to discriminate sensory qualities "usually outstrips one's ability to articulate… the basis of such discriminations in words." Indeed, we can have the concept of "catness" even though we cannot put into words what counts as a cat. We could define "cat" as "a smallish, furry, four-legged predatory mammal with small, sharp teeth, a serpentine tail, a fondness for chasing mice, and a 'meaow'-like cry" [4, pp. 144-145]. Biologists of course would give a more rigorous definition. But we do not need either definition in order to identify a cat as such. "A mute, three-legged feline amputee with a bobbed tail, dull teeth, and all the predatory instincts of a couch pillow will still be reliably identified as a cat by any normal person, even by a child." And by a dog also, we might add.

The brain structures of language grow out of other brain structures. But those underlying structures are already sufficient to account for knowledge (although not for that subset of knowledge which is strictly linguistic, such as knowledge of language).

These considerations extend to scientific and social knowledge, as we will see in the next section.

3. Western Elitism

A very important moment in the development of Western elitism, according to Feyerabend, was the rejection by Socrates of the Homeric worldview. In particular, Socrates would ask his fellow citizens to tell him what virtue, justice, and knowledge were. When they gave him a list of examples in which the word was appropriately used (e.g., the virtue of a man, the virtue of a woman, of a child, etc.), Socrates sarcastically replied that he had asked for one thing and his interlocutor had given him many. Socrates wanted a definition, a universal; they gave him particulars. Greeks, Feyerabend thinks, thought in terms of examples. Indeed, "the view that giving an account means enumerating instances, not subsuming them under a single term, retained its popularity right into the classical age of Greece" [8, p. 38]. Thus, we have two competing models of knowledge: the examples model and the abstraction model. Or, perhaps I should say, we *had* two models, since apparently the examples model was pretty much run over by Western elitism.

I do not believe that it has died, though. In fact, I would wager that it is the way most human beings still think. And there is a good reason for it: That is how the human brain works. Feyerabend's comrade-in-arms, Kuhn, was the first philosopher of science to call our attention to this matter. He argued that, when scientists practice their trade, they do not apply rules but instead learn to see problems as being like other problems they encountered before, where "being like" is best explained by Wittgenstein's notion of "family resemblance" [9]. Scientists are thus trained on a collection of particularly instructive examples ("exemplars") that will enable them to develop a grasp of the way their discipline approaches its investigation of the world. The rules Kuhn would have us do without are the analogs of Socrates' abstract definitions, and thus it was not surprising that his proposal, which expelled from science the sort of decision procedures dear to the hearts of philosophers of science, met with mumblings against Wittgenstenian obscurantism [cf., also 7]. But science has come to Kuhn's rescue.

In the last three decades, scientists and philosophers have collaborated in a program to explain the workings of the human brain (in part) as a complex of neural nets. As Paul Churchland explains, a neural net is designed to compute a large number of functions, even functions that we are *unable to specify*, "so long as we can supply a modestly large set of *examples* of the desired input/output pairs" [3, p. 6]. This process, by the way, is called "training up the network." In artificial networks, the error in the output in the first run is calculated and fed back to the units in the network. This procedure will lead to a readjustment in synaptic weights in the network (this is the "back-propagation" algorithm). After repeating the procedure many times over, the network will finally assume "a configuration of weights that does yield the appropriate outputs for all of the inputs in the training set" [3, p. 7]. We can, for instance, train a network to discriminate sonar echoes of explosive mines from those of submarine rocks, explains Churchland. After it has been trained, the network will be able to identify reliably echoes it has never heard before. It is important to realize that "neural nets typically have no representation of any rules, and they do not achieve their function-computing abilities by following any rules. They simply 'embody' the desired function, as opposed to calculating it" [3, p. 12].

This account sounds very much like Kuhn's explanation of how scientists typically operate. It certainly seems reasonable to consider it a serious model of the typical workings of human neural nets. Contrast it now with the failure of classical artificial intelligence (AI) to explain human thinking in terms of abstract rules.

This elitist philosophical approach extends to morality as well. "Greek morality at the time of Plato," Feyerabend says, "was a morality of instances and examples, not a morality ruled by abstract properties" [8, p. 259]. I would bet that the same could still be said of most fruitful human moralities (as opposed to ethical models invented by philosophers). Churchland has developed this very point in a very provocative way [4, pp. 143-150], [11, pp. 130-147]. For some of the many ways biology may

also influence the evolution and nature of morality (human and animal), the reader may wish to consult [1], [2], [6] and [13].

4. Conclusion

Knowledge need not be linguistic. Moreover, knowledge is the result of adaptive brain structures at work. We can say that an intelligent creature knows because its relevant behavior succeeds. The justified true belief model, therefore, fails to capture characteristic, let alone obligatory, features of knowledge. Since we can dispense with justified true belief as an account of knowledge, we need not concern ourselves unduly with philosophical tricks that seem to confront that account with paradox. Gettier's clever objection becomes a mere curiosity.

References

1. Bernal, J. S. The Role of Sex and Reproduction in the Evolution of Morality and Law, In F. de Sousa and G. Munévar (eds.), *Sex, Reproduction and Darwinism*, London: Pickering & Chatto, 2012, pp. 141-152.
2. Churchland, P. S. *Braintrust: What Neuroscience Tell Us about Morality*, Princeton University Press, 2011.
3. Churchland, P. M. A Deeper Unity: Some Feyerabendian Themes in Neurocomputational Form, In G. Munévar (ed.), *Beyond Reason: Essays on the Philosophy of Paul Feyerabend* (*Boston Studies in the Philosophy of Science, Vol 132*), Boston: Kluwer Academic Press, 1991.
4. _____. *The Engine of Reason, the Seat of the Soul*, Cambridge: Massachusetts Institute of Technology, 1996.
5. Davidson, D. Thought and Talk, In S. Guttenplan (ed)., *Mind and Language*, Oxford: Oxford University Press, 1975, pp. 7-23.
6. De Waal, F. *Primates and Philosophers: How Morality Evolved*, Princeton University Press, 2006.
7. Dreyfus, H. L. *What Computers Can't Do: A Critique of Artificial Reason*, New York: Harper & Row Publishers, 1972.
8. Feyerabend, P. *The Conquest of Abundance: A Tale of Abstraction vs. the Richness of Being*, Chicago: University of Chicago Press, 2000.
9. Kuhn, T. S. *The Structure of Scientific Revolution*, Chicago: University of Chicago Press, 1970.
10. Marcus, R. B. *Modalities: Philosophical Essays*, New York, Oxford: Oxford University Press, 1993.
11. Munévar, G. *Evolution and the Naked Truth: A Darwinian Approach to Philosophy*, Aldershot: Ashgate Publishing Company, 1998.
12. Reber, P. J., M. Beeman, and K. A. Paller. *Human Memory Systems: A Framework for Understanding the Neurocognitive Foundations of Intuition*, International Conference on Augmented Cognition, AC 2013: Foundations of Augmented Cognition, 2013, pp. 474-483.
13. Peterson, D. *The Moral Lives of Animals*, London: Bloomsbury Press, 2011.

sciendo

studia humana
QUARTERLY JOURNAL

Studia Humana
Volume 8:1 (2019), pp. 99—113
DOI: 10.2478/sh-2019-0008

Theodore the Studite's Christology Against Its Logical Background

Basil Lourié

National Research University
Higher School of Economics,
Perm-St. Petersburg, Russia

e-mail: hieromonk@gmail.com

Abstract:
Theodore the Studite resolved the logical problem posed by the second Iconoclasm in an explicitly paraconsistent way, when he applied to Jesus the definition of the human hypostasis while stating that there is no human hypostasis in Jesus. Methodologically he was following, albeit without knowing, Eulogius of Alexandria. He, in turn, was apparently followed by Photius, but in a confused manner.
Keywords: Theodore the Studite, Patriarch Photius, Iconoclasm, Christology, paraconsistent logic.

> Perhaps the most surprising thing, then, is how easily considerations of consistency can be detached from these notions [truth, negation, rationality, and logic], and so how non-integral they are to them. This makes the traditional view of the centrality of consistency to these notions even more surprising. The dead hand of Aristotle has, it would seem, weighed on the topics, preventing philosophers from applying to them the critical spirit which is their due.
>
> Graham Priest [39, pp. 208-209]

1. Introduction: From Under "The Dead Hand of Aristotle"

Dealing with the logic used by the Byzantine patristic authors in their theological reasoning, the modern historians are facing a major problem. At the first glance, they still are in a familiar realm where the Aristotelian logic – whatever the word "Aristotelian" could mean for Byzantium – is not only valid but also considered as *the* logic, that is, the only possible way of sound reasoning. The basic laws of this logic – those of identity, non-contradiction, and excluded middle – are markedly respected. From time to time, however, the steady flow of logical argumentation is interrupted by acceptance of some facts claimed to be "beyond reason and understanding" (ὑπὲρ λόγον καὶ

ἔννοιαν) – to use a formulation from an often repeated at the Byzantine Vespers hymn by John of Damascus.[1] Such facts – also at the first glance only – appear to be illogical at all.

So far, so good. There is *the* logic on the one hand, and there is something "beyond *logos*" on the other. We can preserve such an impression until the moment when we look at the thin interface between the two realms. There, an "Aristotelian" logician, face distorted in horror or distaste, begins to notice a pulsation of some inference, that is, appearance of some conclusions from some premises. The rules of this inference, in general, respect none of the three basic laws of classical logic. Indeed, nobody in the Middle Ages has pretended to include them into the logical textbooks. Nevertheless, the rare thinkers who were attacking them as illegitimate at all (such as John Philoponos[2] or Barlaam the Calabrian,[3] to name only the most known today) were always in danger to be condemned for a heresy. One must confess, however, that many less radical theologians have experienced severe difficulties when they were turned out to admit one or other blatant disruption of the Aristotelian logical laws. As we will see below, among them was even Patriarch Nicephorus of Constantinople.

From a modern point of view, we would prefer to call "logic" anything where there are some procedures of inference, regardless of their particular rules. If the inference is convincing for – or, at least, understandable to – at least, somebody, we can reasonably conclude that the rules of this inference exist. In our modern sense, they also form a logic.

Moreover, there must be a kind of continuity between this n0n-Aristotelian logic and the Aristotelian logic of, say, demonstrative syllogisms that were used in Byzantine theological discussions. Within the Byzantine theological thinking, the Aristotelian "laws" were, indeed, respected, but not on the level of the universal laws *sensu stricto* (there was only one person, in Byzantium, who dared to insist on their applicability even to the divine reality, John Philoponos). Instead, their value was limited to that of the contingent rules of a given domain, namely, the domain of the created.

The proper rules of inference within the interface between the divine domain and the created world could be extracted from the Byzantine theological works and translated into our modern logical language. Here I will propose one case study, that of the Christology of Theodore the Studite.

The unity of the whole system of reasoning in theology was preserved, nevertheless, by the mainstream Byzantine theologians, but not on the level of "laws" or rules but on the level of logical connectives, such as negation or conditional, having the same meaning in all the possible domains of thinking.[4] In general, the most fundamental logical notions which are truth, negation, and rationality were respected throughout the domain of theological reasoning, but the price was logical inconsistency – incompatible with the Aristotelian very notion of thinking.

In the twentieth century, especially since the 1970s, many non-consistent logics are described.[5] These so-called paraconsistent logics made our modern logical thinking ready for grasping logically the meaning of apparently illogical statements of the Byzantine Fathers.

2. Patriarch Nicephorus of Constantinople in the Dead End of the Classical Logic

Throughout the history of the Christian world, there have never been such things as *the* iconoclast theology or *the* theology of icon veneration. On the contrary, there were many different iconoclastic doctrines as well as many different meanings of icon veneration, often incompatible with each other.[6] Fortunately, our present task is limited to a unique and quite specific iconoclastic doctrine as well as a unique and specific kind of theological defence of the holy icons.

At the outbreak of the second iconoclasm (815-842), there was no ready answer to the new version of the iconoclast theology. The iconoclasts managed to show that the current teaching of their opponents is illogical in the pernicious sense, that is, that its logical clarification would lead to either iconoclastic doctrine or Nestorian Christology. This challenge was eventually answered by Theodore the Studite. The logical problem that will be resolved by Theodore the Studite becomes

more understandable against the background of the contemporaneous failed attempt to do the same by Patriarch Nicephorus.

The iconoclasts were perfectly consistent in their demonstration why the icons have nothing to do with the incarnation of the Logos. Their line of thought could be recovered as following:[7]

Starting from

(1) the majority view of the Chalcedonians (shared by the defenders of the icons) that the Logos is incarnated into the common nature of humankind and

(2) a strict conviction shared by all the anti-Nestorians that there is only one hypostasis in Christ, that of the Logos, they have argued, with a reference to

(3) the standard textbook definition of hypostasis (hypostasis = nature + hypostatic features, *idiomata*),[8] that the Logos did not receive the hypostatic features (*idiomata*) of the human personality of Jesus – unless Jesus becomes an additional (human) hypostasis in Christ, beside the Logos.

(4) There is, however, in Christ nothing depictable except these *idiomata* of Jesus – this point was also shared by the defenders of the icons. Therefore, it follows

(5) **the iconoclastic conclusion:** Christ as the incarnated Logos is indepictable, whereas all the human (depictable) features of Jesus are accidental in respect to the incarnated Logos. In other words, the depictable features of Jesus were not those in what the Logos was incarnated and, therefore, are unworthy of any veneration.

According to the ninth-century iconoclasts, Jesus – that is, the conjunction of the personal human features of the incarnated Logos – is *totally* accidental to Logos's incarnation. For the iconophiles, there was no argument about saying that *some* of such human features of Jesus are accidental – those that are accidental to any human (such as the stature or facial expression) – but never those invariant features which make one human individual discernible from all others. For the iconoclasts, however, even those human features that were not accidental to Jesus were accidental to the incarnated Logos.

The defenders of the holy icons shared with their opponents the first four points enumerated above but refused to accept their conclusion as clearly (at least, to them) opposed to the Church Tradition. According to them, something somewhere gone wrong. But where?

Point (2) was certainly out of discussion since the fifth century.

Point (4) was obvious to the two sides of the conflict.

Point (1) has been discussed during the sixth and even the early seventh centuries, but –temporarily – ceased to be under discussion after the victory of the "Maximites" over the Monotheletes;[9] the discussion will be reopened in the eleventh century[10] but not in the ninth.[11]

Point (3) was the weakest point in the whole chain. It has been already dealt with by Maximus the Confessor, but the "Maximites" of this period knew his teaching too superficially to become able to apply it here. Thus, formally, the "school" definition of hypostasis remained unshaken.

The iconoclasts were then, during the second period of Iconoclasm, perfectly fitting with the mainstream theological standard of the epoch but in an apparent conflict with the already ancient custom. Their opponents were in conformity with the custom but without any appropriate theological language at all.

Patriarch Nicephorus was a hostage, if not a victim of the situation of such a theological "mutism."[12] He was able to express his Christology as following: "Nobody of those who have the intellect would accept that either the Logos took the passions or that the flesh undertook the miracles."[13]

This text is not only in contradiction with the "Neo-Chalcedonian" Theopaschism, but even with the Justinianic "Symbol of faith" *Oh Monogenes* (*CPG* 6891), which was then an obligatory part of each Eucharistic liturgy according to the rite that Nicephorus followed himself: "Oh the Only-Begotten Son and the Logos of God… who hast crucified, oh Christ God…"[14]

Moreover, such a Christology contradicts to another part of Christology of the same Nicephorus: he was certainly convicted that the image of Jesus's flesh encompasses the Logos – but he turned out to be unable to explain why.
We see, in Nicephorus, a case when a theological doctrine is completely inadequate to its logical package – a case when *the new wine* of the Orthodox theology runs out from *the old wineskins* of the Aristotelian logic (Lk 5:37). There was an urgent need of new wineskins for preventing the pouring out of the theological wine.

3. The Christology of Theodore the Studite: Its Central Point

A completely new approach has been formulated by Theodore the Studite. There is no direct connexion, as one can see now, between Theodore and the relevant details of the Christology of Maximus the Confessor. In the ninth century, Maxim was still too little known in Byzantium.

Probably, the best description of Theodore's theology as a whole is now provided by Dirk Krausmüller [20]. Therefore, I can go directly to Theodore's main Christological idea.

According to Theodore, the Logos became "one from us" as Jesus – but there was no, in Jesus, a distinct human hypostasis. There was no Jesus as a separate man, but there is Jesus as someone – namely, the divine Logos – having all the features of a separate man, that is, the human nature and the *idiomata* of the separate human hypostasis:[15]

Οὐκ ἄρα μόνῳ τῷ προσηγορικῷ, ἀλλὰ γὰρ καὶ τῷ κυρίῳ ὀνόματι κέκληται ὁ Χριστός· τὸ χωρίζον αὐτὸν τοῖς ὑποστατικοῖς ἰδιώμασιν ἀπὸ τῶν λοιπῶν ἀνθρώπων· καὶ διὰ τοῦτο περιγραπτός. <…>	Therefore, Christ is called not only with a common noun but also with a proper name [*sc.*, Jesus. – *B. L.*] that separates him, *via* the hypostatic features (*idiomata*), from the remaining humans. This is why he is describable. <…>
Οὐκοῦν εἷς ἐστι καθ' ἡμᾶς, εἰ καὶ θεὸς ὁ εἷς τῆς Τριάδος· ὡς ἐκεῖ ἀπὸ τοῦ Πατρὸς καὶ τοῦ Πνεύματος, τῷ υἱκῷ ἰδιώματι διακεκριμένος· οὕτως αὐτὸς ἐνταῦθα ἀπὸ πάντων τῶν ἀνθρώπων τοῖς ὑποστατικοῖς ἰδιώμασιν ἀφοριζόμενος· καὶ διὰ τοῦτο περιγραφόμενος.	Therefore, he is one from us, even though he is God that is one of the Trinity. In the same manner, as he is distinguished there from the Father and the Spirit with the *idiom* of sonship, he is also separated from all the humans here with the hypostatic *idiomata*. And this is why he is describable.

One can feel that Theodore said here something sounding non-Aristotelian. Let us see, however, in more details, what happened here to the three Aristotelian "laws."

4. The Three "Laws" of the Classical Logic in Theodore's Reasoning

4.1. The "Law" of Identity

Aristotle's verbose formulation of the principle of identity in *Metaphysics* IV, 4 implies that anything that could be described in some particular way is always precisely the same thing that can be described in this way.[16] Later Leibniz succinctly put it in a more abstract form: "Ce qui est, est; Chaque chose est ce qu'elle est."[17] This Aristotelian definition of identity through description was further developed into the so-called Leibniz's principle that postulates identity of any two individuals whose all properties are identical. Leibniz himself, during the last months of his life in 1716, acknowledged that "his" principle is not as universal as he himself was arguing shortly before – thus allowing difference between the objects that have absolutely identical properties including

the spatial coordinates (as we see now among the quantum objects such as electrons).[18] This was not, however, compatible with any interpretation of identity that was known to the Antiquity.

Theodore broke the "law" of identity in the following manner. According to his explanation, Jesus is the Logos with no separate human hypostasis. He is not the same as the hypothetical *Jesus that is a human hypostasis (known to Theodore's contemporaries from Nestorian Christology). However, both Jesus and *Jesus have identical properties, that is, the full set of properties of a human individual called Jesus. Both Jesus and *Jesus are unified with the Logos. This feature, though, is to be factored out, in our comparison between the two, because any possible difference in the mode of union between the Logos and the humanity of either Jesus or *Jesus depends exclusively on the possible non-identity between the two.

According to the principle of identity in its standard (Aristotelian) understanding, as well as its explication in the so-called Leibniz's principle, Jesus must be identical to *Jesus – as the iconoclasts would have said in accusing the iconophiles of Nestorianism. Nevertheless, Theodore did not admit this conclusion from the premises he shared with the iconoclasts, because he did not admit the corresponding rule of inference either – which is the rule ("law") of identity. This was a break with the consistent reasoning.

I would add that such a claim was then very risky. Theodore did not know his patristic predecessors who have already dealt with in details the problems of inconsistency of the logic applied to the theological domain. Nevertheless, he certainly imbibed with education the relevant intuitions of Gregory of Nazianzus and Dionysius the Areopagite.

4.2. The "Law" of Non-Contradiction

The principle of non-contradiction is broken by Theodore straightforwardly. According to Theodore, Jesus *is not* a hypostasis of the human nature, but he *is* a human individual in the same manner as everybody of us – "one of us" (εἷς ἐστι καθ' ἡμᾶς, s. above).

Theodore's Jesus is identical with the object that, according to the school definition of hypostasis, is a human hypostasis called Jesus. In the same time, Jesus is not identical with it. Being both identical and not identical to the same thing (namely, the hypostasis of Jesus according to the school definition; we have designed this hypothetical object as *Jesus) is a contradiction.

Both Jesus and *Jesus are identical – in Aristotelian and Leibnizian sense of having identical properties – to the same object, namely, the object of the school definition of hypostasis of the human nature. Indeed, Theodore denied identity between Jesus and *Jesus, but in the way of refusing to call "identity" the relation that is to be called so from a classical (and any consistent logic's!) point of view. According to Theodore, his Jesus is not Nestorian *Jesus only because the identity of properties (*idiomata*) is still not, for Theodore, an identity. As it was to be expected, the breaking of the "law" of identity leaded him to the breaking of the "law" of non-contradiction (or *vice versa*).

Thus, in classical (and not Theodore's) terms, we obtain a subcontrary (not contradictory nor contrary) opposition: Jesus is identical to *Jesus, whereas it is claimed, by Theodore, that he is not.

In classical terms, this means that $A = B$ but $A \neq B$ simultaneously.

The principles of identity and non-contradiction are so mutually depending that there is no possibility of breaking one without breaking another.

Let us explain Theodore's intuition in a more Aristotelian fashion, using [different variables's values for the same functions, that is] examples of a human and a horse, so dear to the antique philosophers. Then, Theodore's reasoning could be reformulated as following. Some individual (hypostasis) has, for instance, the features (*idiomata*) of both human Peter and horse Pegasus; however, this hypostasis has these features not partially, as a centaur, but of both of them entirely. He is entirely Peter and entirely Pegasus. Even though he is, among the horses, a horse called Pegasus, he is still a human among the humans whose name is Peter.

For a viewpoint of any logic respecting the law of non-contradiction, such a claim is impossible. Instead, such logic would allow only two kinds of compositions: (1) some mixed cases,

such as some hybrid, centaur, resulted from Peter and Pegasus, which is no longer identical to Peter or Pegasus, or (2) a two-individual set formed by Peter and Pegasus taken together as two different elements of the one set. One can easily recognise, in the first alternative, the decision of the Monophysitism, and, in the second alternative, the decision of the Nestorianism.
The first alternative is, from a historical point of view, even more interesting, even though it was not mentioned in the discussions of the ninth century. It is quite important for understanding the origin of the logical problems in Christology that Theodore was facing. As one could guess, it concerns the principle of the excluded middle and breaking thereof.

4.3. The "Law" of Excluded Middle

Some limitations of the "law" of excluded middle were known to Aristotle and other antique logicians who have described the modes of reasoning which we now call modal. Aristotle himself described the first of the known modal logics now called alethic, where he used such categories as "necessarily", "impossibly", and "possibly" instead of the bivalent statements "true" or "false". The alethic modal logic is perfectly Aristotelian, too, but not classical. Thus, it was known to the antique logicians that principle of the excluded middle is not obligatory for making reasoning consistent.[19]

Ironically, among the three "laws" of the classical logic, this one is the only one that Theodore respects. To him, there is nothing in between of Jesus and *Jesus: the real Jesus could be either a hypostasis of the Holy Trinity (Jesus) or a hypostasis of the human nature (*Jesus) but never something third. The latter possibility is excluded *a priori*, whereas the second one (that Jesus is a human hypostasis) only *a posteriori*, as a conclusion of Theodore's theological analysis. This manner of thinking is in the perfect accordance with the principle of excluded middle in a completely consistent and even classical way, albeit Theodore's claim that Jesus has all properties (*idiomata*) of *Jesus without being *Jesus is breaking the consistence of reasoning.

Instead of looking for a *tertium quid* between Jesus and *Jesus, Theodore appropriated *Jesus's features to Jesus in a paraconsistent way. In consistent terms, we have already described this procedure as simultaneous identification and non-identification between the two. Such an operation requires that the binary opposition between Jesus and *Jesus is duly respected and nothing in between of them is introduced.

Let us consider another hypothetical situation, when we need to preserve the consistence of reasoning but also to avoid Nestorianism. This is the situation when some consistent *tertium quid* between the Nestorian *Jesus and Theodore's paraconsistent Jesus becomes necessary. This would mean that the divine hypostasis of the Logos, after having become composite with acquiring humanity, formed as well a nature of its own, μία φύσις τοῦ θεοῦ λόγου σεσαρκομένη ("the one nature of the God Logos incarnated") – in some of the meanings of this extremely multivalued expression.

In the consistent reasoning, the Logos could never become a hypostasis of the human nature. If, nevertheless, he accepted Jesus without accepting a separate human hypostasis (that is, without accepting the Nestorian *Jesus), then, the Logos and Jesus are now the same hypostasis. In Theodore's paraconsistent reasoning, the hypostasis of the Logos and Jesus is also the same, but "Jesus" became the name of the Logos according to the human nature – in a paraconsistent way. In our present hypothetical situation, any paraconsistent way is forbidden. Thus, the Logos does not have a name according to the human nature, because he did not become a hypostasis of this nature either. However, "Jesus" is not a name of something belonging to the divine nature – which is obvious unless we accept the extremist Christology of the so-called actistism.[20] Therefore, the object fitting with "Jesus" as its name must be defined as a new separate nature, distinct from the natures of humanity and divinity.

Our hypothetical situation, of course, took place in the history. This is the reasoning by John Philoponos shortly before the Fifth Ecumenical Council (553), when he interpreted "the unique hypostasis" of Christ in the Chalcedonian sense as identical with the "unique nature" of Christ of the n0n-Chalcedonians.[21] This was an anti-Nestorian and completely consistent decision. The

Chalcedonians, in turn, were ready to acknowledge in μία φύσις of Cyril of Alexandria the Chalcedonian "unique hypostasis" but did not agree with this Philoponian reverse moving asking them interpreting their own "unique hypostasis" as the anti-Chalcedonian "unique nature."
The Christology of the second Iconoclasm was also anti-Nestorian and completely consistent, but Philoponos would dislike it for almost the same reasons as the ninth-century iconophiles. For both Philoponos and the iconophiles, the iconoclastic negation of the individual humanity in the incarnation of the Logos would look equal to denying the reality of the incarnation and, therefore, a kind of "phantasiasm," according to the heresiological jargon of the epoch.

Both Theodore and his iconoclast opponents were anti-Monophysite in the sense that all of them denied the Philoponian identification of "hypostasis" and "nature" in Christ. Such a "unique nature" would be a *tertium quid* between the paraconsistent Jesus of Theodore and the Nestorian *Jesus.

5. Patriarch Eulogius of Alexandria, a Theoretician of Paraconsistency

In the epoch opened with the Triumph of Orthodoxy in 843, Photius was the person who undertook a revalorisation of the theological legacy available to him. Maximus the Confessor, as it seems remained mostly beyond his horizon. He became very successful, however, in collecting the works of the authors of the sixth century.[22] Patriarch Eulogius of Alexandria (580-607) was among them.

The sixth among his eleven treatises summarised by Photius was written on the 580s discussion between the Severianist patriarchs of Alexandria and Antioch, Damian and Peter respectively, and especially against the position of Peter. Thus, this treatise was aimed "against verbiage of those who consider the hypostasis to be only an idioma (ἰδίωμα μόνον)." Damian would look an easy sparring-partner, in such an extent his attitude was at odds with the Cappadocian Fathers.[23] Nevertheless, in fact, it was not so. The problem was in the search of an alternative to the Damianism, which failed to provide his opponent Peter.

According to Eulogius,[24] both opponents were not right. They both misunderstood the meaning of the definition of hypostasis that they quote – for instance, from Basil of Caesarea. Indeed, Basil has said that the hypostasis is a superposition of the nature/essence and the idiomata. This definition, indeed, implies some complexity and, therefore, contradicts to the absolute simplicity of God. Nevertheless, this complexity is strictly limited to the capacity of our mind, whereas there is no complexity in God.

Φασὶ γάρ τινες συμπλοκὴν οὐσίας καὶ ἰδιώματος εἶναι τὴν ὑπόστασιν· ὃ περιφανῶς συνεισάγειν οἶδε τὴν σύνθεσιν, καὶ ποῦ ἂν εἴη τὸ ἁπλοῦν καὶ ἀσύνθετον τῆς ἐν τῇ Τριάδι Θεότητος; Οἱ δὲ καὶ Βασίλειον προϊστῶσι τὸν μέγαν τῆς φωνῆς διδάσκαλον, οὐκ ἐθέλοντες νοεῖν ὡς ὁ σοφὸς ἐκεῖνος ἀνὴρ οὔτε ὅρον οὔτε ὑπογραφὴν ἀποδιδοὺς ὑποστάσεως τὸ τῆς συμπλοκῆς παρέλαβεν ὄνομα, ἀλλὰ βουλόμενος ἐπιστομίσαι τὸν ἀνόμοιον τὴν ἀγεννησίαν καὶ τὴν οὐσίαν εἰς ταὐτὸν ἀγαγεῖν φιλονεικήσαντα, καὶ τὴν πρὸς τὸ γεννητὸν τοῦ ἀγεννήτου διαφορὰν εἰς τὸν τῆς οὐσίας λόγον μεταγαγεῖν, ἵνα μὴ μόνον διαφόρους, ἀλλὰ καὶ ἀντικειμένας οὐσίας εἰσάγοι ἐπί τε τοῦ Πατρὸς καὶ τοῦ Υἱοῦ.	Certains disent en effet que l'hypostase est l'union [conjunction] d'essence et de propriété, proposition nettement susceptible d'amener la notion de composition ; et d'où serait le caractère simple et exempt de composition de la Trinité divine ? Ces gens-là vont même jusqu'à mettre en avant Basile le Grand qui aurait enseigné cette formule et ils ne veulent pas comprendre que ce grand sage n'a pas défini ou décrit l'hypostase quand il a employé le mot union [conjunction] mais qu'il voulait imposer silence à l'Anoméen qui prétendait réduire à l'identité l'incréé et l'essence et ramener la différence entre le créé et l'incréé à l'idée d'essence pour aboutir, à propos du Père et du Fils, à l'idée d'essences non seulement différentes mais opposées.
Διὰ τοῦτο *ὁ τοὺς λόγους οἰκονομῶν ἐν κρίσει*	C'est pourquoi Basile, qui *règle ses paroles en conscience*, dans sa discussion avec

[Ps 111:5] Βασίλειος, ἐν τῷ πρὸς τὸν ἀνόμοιον ἀγῶνι, τῷ κοινῷ συμπλέκει τὸ ἴδιον, ἀσύγχυτον ἡμῖν καὶ διακεκριμένην μεθοδεύων τὴν τῆς ἀληθείας κατάληψιν. Ἀπορεῖ μὲν γὰρ ὁ ἀνθρώπινος νοῦς ἁπλῇ καὶ μιᾷ προσβολῇ τὸ ἑνιαῖον ἅμα καὶ ἁπλοῦν καὶ τὸ τρισσὸν καταλαβεῖν τῶν ὑποστάσεων· διὸ τῇ τῶν ἰδιωμάτων, ὡς ὁ διδάσκαλος ἔφη, προσθήκῃ τὴν ἰδιάζουσαν ἀφορίζει τῶν ὑποστάσεων ἔννοιαν καὶ ἔστι μὲν ἡ μέθοδος ἀσθενείας ἐπίκουρος καὶ τῆς περὶ τὸ ἀκατάληπτον συνεργὸς καταλήψεως, οὐ μήν γε συμπεπλεγμένον τὸ ἁπλοῦν τῆς θεότητος ἢ ὅλως τινὰ τῶν ταύτης ὑποστάσεων οὐμενοῦν οὐδαμῶς ἀπεργάσαιτο. Διὸ καὶ ἐπήγαγεν ὡς ἀμήχανον ἰδιάζουσαν ἔννοιαν Πατρὸς λαβεῖν ἢ Υἱοῦ, μὴ τῇ τῶν ἰδιωμάτων προσθήκῃ τῆς διανοίας διαρθρουμένης. Καὶ ὅπερ ἐν τοῖς προλαβοῦσι συμπλοκὴν ἐκάλεσε, τοῦτο νῦν προσθήκην ὠνόμασε. Σαφέστερον δὲ τὸ εἰρημένον ποιῶν· «Οὐ γὰρ οἱ δεικτικοί, φησί, τῆς ἰδιότητος αὐτοῦ τρόποι τὸν τῆς ἁπλότητος αὐτοῦ λόγον παραλυπήσουσιν· ἢ οὕτω γε ἂν πάντα, ὅσα περὶ Θεοῦ λέγεται, σύνθετον ἡμῖν τὸν Θεὸν ἀναδείξῃ»[25].	l'Anoméen, unit le particulier au commun en nous montrant comment comprendre la vérité sans confusion et dans une clarté absolue. L'esprit humain et en effet embarrassé quand il s'agit de saisir d'un simple et unique mouvement [grasping – *B.L.*] les notions d'unité et de simplicité en même temps que celle des trois hypostases. C'est pourquoi, comme l'enseigne le maître, c'est par l'addition des propriétés qu'il détermine sa propre conception des hypostases, et cette façon de procéder est un secours pour la faiblesse et une aide pour comprendre l'incompréhensible, mais Basile ne transformait absolument pas en un composé la simplicité de la divinité ni, en un mot, aucune des hypostases divines. C'est pourquoi il a ajouté qu'il est impossible de se faire une conception propre du Père et du Fils sans que notre pensée se complète par l'addition des notions de propriété ; et ce qu'il avait auparavant nommé union [conjunction – *B.L.*], il l'appelle maintenant addition. Et pour rendre sa parole plus claire : « Ce ne sont pas, dit-il, les façons de montrer ses caractères spécifiques qui nuiront à sa façon d'envisager la simplicité ; sinon, tout ce qu'on nous dit de Dieu démontrerait que Dieu est un composé ».

Let us ask Eulogius: Ok, there is no complexity in God, whereas the hypostasis is, by definition, something complex. Then, how you insist that there are hypostases in God at all?

For Eulogius, however – as well as for Peter and Damian – the presence of three hypostases in the unique God was out of question. This was simply a received knowledge.

Thus, Eulogius repeats the "school" definition of hypostasis but adds that, in God, there is no room for hypostases, whereas hypostases themselves there are. A hypostasis in God is something that is impossible in God but that is.

Then, one can approach this problem from the opposite side asking Eulogius: Why do you call these logical objects in God "hypostases," if you acknowledge that the hypostasis is, by definition, something else than anything that could occur in God? For answering, Eulogius would refer to an established patristic tradition that could be called "The Correspondence Principle."

6. The Correspondence Principle

Today it became easy to answer such questions. We are in presence of a just another instance of applicability of the principle that Niels Bohr called the Correspondence Principle. In Bohr's Copenhagen interpretation of the Quantum theory, this means that the notions of classical physics continue to be used for description of the quantum reality but in a non-classical way.[26] In the same manner, in Eulogius's explanation, the "classical" definition of hypostasis and the notion of hypostasis itself continues to be used, but not in a "classical" Aristotelian way. In both cases, in Bohr's Quantum theory and Eulogius's Triadology, the "classical" notions change their meaning,

and, in both cases, there is no direct way to make these changes explicit – except an indirect way that is actually used.

The notion of hypostasis applied to God is no more classical than the notion of spatial coordinates applied to an electron.

The Quantum logics proposed for the Copenhagen interpretation of the Quantum theory, especially in the 1990s and later, are inconsistent.[27]

Now we can say that Eulogius of Alexandria has explained that the logic used by the Cappadocian Fathers was, in fact, a paraconsistent one. Let us emphasise an important thing: Eulogius has never said that some classical notions are applied to God in an approximate way and not in the proper sense. He says exactly the opposite: they are applied in their proper and exact sense. However, they are inapplicable. The theological meaning is contained not in the simply procedure of application of some notions to God but in a double procedure of such application joined with insisting of their inapplicability. This conjunction of application and inapplicability forms the difference between the paraconsistent usage of the categories of consistent logic and their approximate usage in a somewhat metaphorical sense.

7. The Paraconsistent Logical Core of Theodore the Studite's Christology

The Christological model of Theodore the Studite is derived from the teaching of Gregory of Nazianzus and Maximus the Confessor on the deification of the man.[28] This teaching implies a logical model often called by modern historians *tantum-quantum* (τοσοῦτον-ὅσον): in as much as the Logos became the man, in the same extent the man – any deified man – will become God, and this extent is, of course, "completely." Nevertheless, the deified man does not become a new hypostasis of the divine nature – as well as the Logos did not become a new hypostasis of the human nature.

In fact, Theodore the Studite's Christology was already present in Maximus the Confessor. There were some differences, however. On the one hand, Theodore made explicit some ideas of Maximus: his Christology is in the mirror symmetry towards Maximus's doctrine of deification. On the other hand, Theodore has never elaborated on Maximus's sophisticated concept of τρόπος ὑπάρξεως.

Let us reformulate the main logical notions used in Christology in a more analytical language. We will use a language of a "set theory," but not of one of the set theories presently used in mathematics but of a kind of "naïve" set theory, closer to its original form in Georg Cantor – where all the paradoxes are tolerated, and there is no difference between the notions of set and class.

Then, the notion of hypostatic *idiomata* becomes equivalent to the notion of being a given element of a class. The *idiomata* feature an individual as a specific individual, whereas not within an unordered universe but within a definite class. Thus, the *idioma* of sonship ("to be begotten") is featuring the Son within the divine nature only, whereas within other natures the notion of sonship does not exist in the same sense. Thus, it is important to note that the hypostatic features do not define a specific individual of whatever nature but only an individual of a given nature, that is, within a given class.

Thus, we can write, for an individual x_i, that is, for the i-th individual of the class X, that to have the *idiomata* of x_i, means that x_i belongs to X, $x_i \in X$.

This definition could be easily applied, in slightly modified forms, to the classes whose elements are uncountable or countable in some inconsistent manner only. An example of such class is the class of hypostases of the Holy Trinity. An approximation of this class with a well-ordered set (that as, a set for which exists a bijection between all the elements of this set and the set of natural numbers) would be a source of misunderstandings or errors in triadological reasoning, because if the Trinity is a set, then, this set is not a well-ordered one nor ordered at all.[29] For the further, however, we need only a very weaken conception of ordering: in this sense, "ordered" is every class where the elements could be discerned in whatever way. In this weaken sense, the class of the

hypostases of the Holy Trinity is, indeed, ordered. Therefore, our (weaken) conception of "being the *i*-th element of a class" is applicable here too, presuming that *i* here is not a natural number and not necessarily a consistent number.[30]

Now, let us consider the case of the incarnated Logos. Without ceasing to be an element of the class "divine nature", he becomes an element of the class of humans when he takes the human *idiomata* of Jesus. Nevertheless, he does not become an element of the class of humans because the Logos does not become a human hypostasis. Therefore, Logos's inclusion into the class of humans is paraconsistent only, whereas his inclusion into the class "divine nature" is consistent: the Logos is a divine hypostasis, and there is no sense in what the Logos ceased to be a divine hypostasis. Thus, the Logos became a human individual called Jesus in a paraconsistent way only.

In a symmetrical way, we have to understand Maximus's (and Gregory of Nazianzus's) doctrine of deification. A human person Peter continues to be a human person in a consistent way but becomes God (the only God in whom the Christians believe) in a paraconsistent way.

8. The Photian Epilogue

As a historian of Byzantine dogmatic discussions can feel, the paraconsistent claims of one or other outstanding Byzantine theologian have required too much intellectual stress for their adequate adaptation by the official theological mainstream. The philosophical culture of the Byzantine court theologians, *de facto* secular, was one of the main obstacles. This is an important reason why the Byzantine *Dogmengeschichte* was not anyhow smooth but highly turbulent and, to say properly, cyclic. The bright logical ideas have been quickly fossilised within the official "repetitive theology," with an inevitable effect of a new confusion that provoked, in turn, reordering and correcting based on new insights of other bearers of bright logical intuitions for theology. Then, a new cycle has begun.

Any paraconsistent theological claim put into the framework of the "repetitive theology" is fossilised in the same way as a poem paraphrased in prose or a joke "explained" to those who have no sense of humour. What remains after such "repetitions" is not the genuine theological meaning that certainly has evaporated.

The Christological ideas of Theodore the Studite did not escape the common destiny, that is, fossilisation and confusion. The references to the Studite by both sides of the quarrels on the holy icons in the late eleventh-century Byzantium form a sufficient proof of this.[31] It is interesting, however, to trace the reception history of Studite's Christology in earlier times.

Patriarch Photius, writing between 867 and 877, repeated Theodore's Christological thesis when answering to a – imaginary or not – iconoclast opponent.[32] The opponent seemed to push Photius toward iconoclasm starting from Patriarch's expected rejection of the Nestorian idea of the incarnation into a particular man. He then put before Photius an alternative: the Logos incarnated into either particular man (τὸν ἐπὶ μέρους [ἄνθρωπον]) or the man in general (τὸν καθόλου ἄνθρωπον) [48, pp. 14-15]. Photius's answer is "Neither". Following the Studite, he wrote: "We say that even if he [the Logos] assumed human nature, the Logos exhibits [its] features (*idiomata*) as his own."[33] Certainly, Theodore's thesis is "repeated" – in the sense of "repetitive theology," at least.

Nevertheless, a new problem arose, and, so far, it remains unknown in what extent Photius resolved it or, at least, realised it.[34] In the present answer, Photius failed to provide an explicit treatment of distinction between the notions of "the man in general" and "the human nature." As the first step, he follows an argument of the earlier iconophilic theologians stating that "the man in general" that is not instantiated in any particular human individual is an abstraction without any real content and, therefore, is incompatible with the reality of the incarnation.[35]

Then, however, he uses against his opponent a classical argument of the anti-Chalcedonians against the Chalcedonians, known since, at least, *ca* 519, the discussion between Severus of Antioch and the Chalcedonian Sergius the Grammarian:[36] the common is to be seen in plurality of hypostases; thus, if Christ is "the man in general", he must have many hypostases, viz. those of the whole human genus.[37]

This argument is at odds with the previous one. If "the man in general" is a mere abstraction, as it has been just stated, it contains no hypostases at all, but if "the man in general" is to be instantiated in plurality of hypostases, it is not a mere abstraction. If it is not a mere abstraction – which was, in Byzantium, the majority opinion – one would like to know what is the difference between "the man in general" in this sense and the human nature assumed by the Logos. Photius failed to provide any explanation. He confused different understandings of *universalia* and, apparently without knowing, repeated a standard anti-Chalcedonian argument. In this way, his argumentation was in a mirror symmetry with Nicephorus's verbal "Nestorianism."

Photius did not look for a recourse to the paraconsistency. Instead, he added two more arguments – demonstrating in what extent the very idea of logical paraconsistency was repulsive to him. The following two questions [48, p. 15.11-19] go immediately after the argument we have just quoted:

συμβαίνει δὲ καὶ μὴ εἶναι ἡμῖν αὐτὸν ὁμοούσιον· ἀναληφθέντος γὰρ τοῦ καθόλου ἀνθρώπου ἐν τῷ θεῷ λόγῳ οὐκέτι ἡμεῖς ἄνθρωποι λεχθείημεν· πόθεν γὰρ τοῦτο ὑπάρξει ἡμῖν, καὶ κατὰ τί κοινωνήσομεν τῷ Χριστῷ ;	Moreover, it follows [from the supposition of the incarnation into the man in general] that he [the Logos] is not consubstantial to us. Indeed, if the man in general is assumed into the God Logos, we are no longer to be called men. On what ground he will be so [*sc.*, consubstantial] to us, and in what sense we will have communion with Christ?
πρὸς δ' αὖ τοῖς εἰρημένοις καὶ ἕτερον ὑπάρξει ἀτόπημα, τὸ τῶν ἀνθρώπων ἕκαστον καὶ ἄνθρωπον εἶναι καὶ μὴ ἄνθρωπον· ἕκαστος γὰρ ἡμῶν κατὰ τὸν ἀληθῆ λόγον ἄνθρωπός ἐστί τε καὶ ὀνομάζεται· τοῦ δὲ καθόλου, καθ' ὃ πάντες ἄνθρωποι εἶναι ἐλέγοντο, παρὰ τοῦ θεοῦ λόγου ἀναληφθέντος, πῶς ἐσόμεθα ἄνθρωποι ;	Moreover, in addition to the already said, there will be another absurdity: everybody from the men would be man and not man. Everybody from us is in the true sense man and is [so] called. However, if the general, according to which all [men] are called to be men, is assumed [*sc.*, withdrawn] by the God Logos, how we will be men?

In both questions, the humanity supposedly assumed by the Logos is taken as different from our humanity – without becoming, however, a humanity of an individual human being. If the universal humanity is assumed by the Logos, it becomes withdrawn from us. Photius showed a clear intuition that the universal humanity could not be shared with us by the Logos in any consistent way, and, therefore, he provided his example of bad inconsistency where we are both to be and to be not men. Photius, thus, tried to avoid dealing with the general humanity in his Incarnation doctrine and, instead, explained the Incarnation as assuming of the human *idiomata* by the Logos. Nevertheless, he had no option to stop calling this fact "assuming of the human *nature*." Then, what means, in this text of Photius, the notion "nature", if he clearly distinguished it from the general humanity (τοῦ καθόλου), "according to which all [men] are called to be men"? I guess that this problem has never been thought through by Photius.

This example of Photius shows that if you throw paraconsistency out of the door, it will come back through the window – in this case, through a confused usage of the terms for universals.

9. Conclusion

Theodore the Studite has been forced to explain why the normal rule of superposition of the classical categories, φύσις + ἰδιώματα = ὑπόστασις, does not work in the case of Jesus: not because this rule is erroneous but exactly because it is correct. Its correctness becomes forceless, thus showing the paraconsistent logic of the divine incarnation overcoming the consistency of human rationality.

If Jesus's human features are not accidental, despite what the iconoclasts claimed, the only remaining solution within the framework of the "Neo-Chalcedonian" Christology is paraconsistent.

References

1. Arenhart, J. R. B., and D. Krause. Potentiality and Contradiction in Quantum Mechanics, In A. Koslow and A. Buchsbaum (eds.), *The Road to Universal Logic. Festschrift for the 50th Birthday of Jean-Yves Béziau*, vol. 2, Basel: Birkhäuser, 2015, pp. 201-211.
2. Baranov, V., and B. Lourié. The Role of Christ's Soul-Mediator in the Iconoclastic Christology, In G. Heidl and R. Somos (eds.), *Origeniana nona: Origen and the Religious Practice of His Time. Papers of the 9th International Origen Congress, Pécs, Hungary, 29 August-2 September 2005*, (Bibliotheca Ephemeridum theologicarum Lovaniensium, 228), Leuven Walpole, MA: Peeters, 2009, pp. 403-411.
3. Baranov, V. 'Condensing and Shaping the Flesh…': The Incarnation and the Instrumental Function of the Soul of Christ in the Iconoclastic Christology, In S. Kaczmarek, H. Pietras, and A. Dziadowiec (eds.), *Origeniana decima: Origen as Writer. Papers of the 10th International Origen Congress, University School of Philosophy and Education "Ignatianum", Kraków, Poland, 31 August-4 September 2009*, (Bibliotheca Ephemeridum theologicarum Lovaniensium, 244), Leuven Walpole, MA: Peeters, 2011, pp. 919-932.
4. Baranov, V. *Amphilochia* 231 of Patriarch Photius as a Possible Source on the Christology of Byzantine Iconoclasts, *Studia Patristica* 68, 2013, pp. 371-380.
5. Barnes, J. (ed.). *The Complete Works of Aristotle*, The Revised Oxford Translation, vol. 2, (Bollingen series 71:2), Princeton: Princeton University Press, 1984.
6. Basil of Caesarea. *Adversus Eunomium libri V*, *PG* 29, 497-768.
7. Beall, J. C., and G. Restall. *Logical Pluralism*, Oxford: Clarendon Press, 2006.
8. Chernoff, F. Leibniz's Principle of the Identity of Indiscernibles, *The Philosophical Quarterly* 31, 1981, pp. 126-138.
9. da Costa, N. C. A., and D. Krause. The Logic of Complementarity, In J. van Benthem et al. (eds.), *The Age of Alternative Logics: Assessing Philosophy of Logic and Mathematics Today* (Logic, Epistemology, and the Unity of Science, 3), Dordrecht: Springer, 2006, pp. 103-120.
10. Diekamp, F. *Doctrina Patrum de incarnatione Verbi. Ein griechisches Florilegium aus der Wende des 7. und 8. Jh.* 2, Aufl. mit Korr. und Nachträgen von B. Phanurgakis, hrsg. von E. Chrysos, Münster: Aschendorff, 1981.
11. Ebied, R. Y., A. Van Roey, and L. R. Wickham (eds.). Peter of Callinicum. *Anti–Tritheist Dossier (*Orientalia Lovaniensia Analecta, 10), Leuven: Peeters, 1981.
12. Follieri, H. *Initia Hymnorum Ecclesiae Graecae*, vol. 2, Città del Vaticano: Biblioteca Apostolica Vaticana, 1961.
13. French, S., and D. Krause. *Identity in Physics: A Historical, Philosophical, and Formal Analysis*, Oxford: Clarendon Press, 2006.
14. Henry, R. (ed.). Photius, *Bibliothèque*. Tome V *(« Codices » 230–241)*, (Collection byzantine), Paris: Les Belles Lettres, 1967.
15. Konstantinovsky, J. Dionysius the Areopagite versus Aristotle? The Two Points of Reference for Gregory Palamas in Initial Confrontations with Barlaam the Calabrian, *Studia Patristica* 42, 2006, pp. 313-319.
16. Krausmüller, D. An embattled charismatic: Assertiveness and invective in Nicetas Stethatos' Spiritual Centuries, *Byzantine and Modern Greek Studies* (forthcoming).
17. Krausmüller, D. Between Tritheism and Sabellianism: Trinitarian Speculation in John Italos' and Nicetas Stethatos' Confessions of Faith, *Scrinium* 12, 2016, pp. 261-280.
18. Krausmüller, D. Divine Genus – Divine Species: John Philoponus' Impact on Contemporary Chalcedonian Theology, In J. E. Rutherford (ed.), *The Mystery of Christ in the Fathers of the Church: Essays in honour of D. Vincent Twomey SVD*, Dublin: Four Courts Press, 2011, pp. 94-105.

19. Krausmüller, D. Hiding in Plain Sight: Heterodox Trinitarian Speculation in the Writings of Nicetas Stethatos, *Scrinium* 9, 2013, pp. 255-284.
20. Krausmüller, D. On the relation between the Late Antique and Byzantine Christological discourses: Observations about Theodore of Stoudios.' *Third Antirrheticus. Jahrbuch der Österreichischen Byzantinistik* (forthcoming).
21. Krausmüller, D. Responding to John Philoponus: Hypostases, Particular Substances and *Perichoresis* in the Trinity. *Journal for Late Antique Religion and Culture* 9, 2015, pp. 13-28.
22. Lang, U. M. *John Philoponus and the Controversies over Chalcedon in the Sixth Century. A Study and Translation of the "Arbiter"* (Spicilegium Sacrum Lovaniense, 47), Leuven: Peeters, 2001.
23. Larchet, J.-Cl. *La divinisation de l'homme selon saint Maxime le Confesseur* (Cogitatio Fidei, 194), Paris: Cerf, 1996.
24. Lebon, I. (ed.), Severi Antiocheni *Liber contra impium Grammaticum*, *Oratio prima et secunda* (CSCO, vols. 111-112; Scriptores Syri, tt. 58-59) [Ser. IV, t. IV], Paris: E typographeo republicae, 1938.
25. Leibniz, G. W. *Sämtliche Schriften und Briefe*, Hrsg. Akademie der Wissenschaften der DDR, Reihe 6., Bd. 6. Berilin: Akademie–Verl., 1990.[2]
26. Lourié, B. A Logical Scheme and Paraconsistent Topological Separation in Byzantium: Inter–Trinitarian Relations according to Hieromonk Hierotheos and Joseph Bryennios, In D. Bertini and D. Migliorini (eds.), *Relations: Ontology and Philosophy of Religion* (Mimesis International. Philosophy, n. 24), [Sesto San Giovanni (Milano)]: Mimesis International, 2018, pp. 283-299.
27. Lourié, B. Damian of Alexandria, In S. Uhlig (ed.), *Encyclopaedia Aethiopica*, vol. 2, Wiesbaden: Harrassowitz Verlag, 2005, pp. 77-78.
28. Lourié, B. Julianism, In S. Uhlig (ed.), *Encyclopaedia Aethiopica*, vol. 3, Wiesbaden: Harrassowitz Verlag, 2007, pp. 308-310.
29. Lourié, B. Le second iconoclasme en recherche de la vraie doctrine. *Studia Patristica* 34, 2001, pp. 145-169.
30. Lourié, B. Leontius of Byzantium and His 'Theory of Graphs' against John Philoponus" In M. Knežević (ed.), *The Ways of Byzantine Philosophy* (Contemporary Christian thought series, no. 32), Alhambra, CA: Sebastian Press, Western American Diocese of the Serbian Orthodox Church; Faculty of Philosophy, Kosovska Mitrovica, 2015, pp. 143-170.
31. Lourié, B. Nicephorus Blemmydes on the Holy Trinity and the Paraconsistent Notion of Numbers: A Logical Analysis of a Byzantine Approach to the *Filioque*, *Studia Humana* 5, 2016, pp. 40-54.
32. Lourié, B. Philosophy of Dionysius the Areopagite: Modal Ontology, In A. Schumann (ed.), *Logic in Orthodox-Christian Thought*, Heusenstamm bei Frankfurt: Ontos-Verlag, 2013, pp. 230-257.
33. Lourié, B. Une dispute sans justes: Léon de Chalcédoine, Eustrate de Nicée et la troisième querelle sur les images sacrées, *Studia Patristica* 42, 2006, pp. 321-339.
34. Lourié, B. What Means "Tri–" in "Trinity"? An Eastern Patristic Approach to the 'Quasi-Ordinals,' *Journal of Applied Logic* (forthcoming).
35. [Lourié, B.] Лурье, В. М. *История византийской философии. Формативный период* [*A History of the Byzantine Philosophy: The Formative Period*], Санкт-Петербург: Axiōma, 2006.
36. Nicephorus of Constantinople, *Refutatio et eversio deliramentorum… [Antirrhetici adversus Constantinum Copronymum]*, In *PG* 100, 205-533.
37. Patterson, R. *Aristotle's modal logic. Essence and entailment in the* Organon, Cambridge: Cambridge University Press, 1995.
38. Priest, G., K. Tanaka, and Z. Weber. Paraconsistent Logic, In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition),
https://plato.stanford.edu/archives/sum2018/entries/logic-paraconsistent/.
39. Priest, G. *Doubt Truth to Be a Liar*, Oxford: Oxford University Press, 2006.

40. Rosenfeld, L. (ed.). *Niels Bohr. Collected Works*, vol. 3: *The Correspondence Principle (1918-1923)*, J. Rud Nielsen (ed.), Amsterdam New York Oxford: North-Holland Publ. Co., 1976.
41. Sinkewicz, R. E. A New Interpretation for the First Episode in the Controversy between Barlaam the Calabrian and Gregory Palamas, *Journal of Theological Studies* 31, 1980, pp. 489-500.
42. Sinkewicz, R. E. The Solutions Addressed to Georges Lapithes by Barlaam the Calabrian and Their Philosophical Context, *Medieval Studies* 43, 1981, pp. 151-217.
43. Tanaka, K., F. Berto, E. Mares, and F. Paoli (eds.). *Paraconsistency: Logic and Applications* (Logic, Epistemology, and the Unity of Science, 26), Dordrecht: Springer, 2013.
44. Theodore the Studite. *Antirrhetici adversus iconomachos*, In *PG* 99, 327-436.
45. Treadgold, W. T. *The Nature of the Bibliotheca of Photius* (Dumbarton Oaks Studies, 18), Washington, D C: Dumbarton Oaks Center for Byzantine Studies, 1980.
46. Van Roey, A. Les fragments trithéites de Jean Philopon, *Orientalia Lovaniensia Periodica*, vol. 11, 1980, pp. 135-163.
47. Wellesz, E. *A History of the Byzantine Music and Hymnography*, 2nd ed., Oxford: Clarendon Press, 1961.
48. Westerink, L. G. (ed.). *Photii Patriarchae Constantinopolitani Epistulae et Amphilochia*, vol. 6, fasc. 1, (Bibliotheca scriptorium graecorum et romanorum Teubneriana), Leipzig: Teubner, 1987.
49. Васюков, В. Л. Логика Галена: наследие Аристотеля или научная инновация? [Vasyukov, V. L. Galen's Logic: Aristotelian Heritage or Scientific Innovation?], *История медицины*, vol. 2, 2015, pp. 5-16.
50. Желтов, М. "Догматик" [Zheltov, Mikhail. "Dogmatikon"] In *Православная энциклопедия*, vol. 15, Москва: ЦНТ «Православная энциклопедия», 2007, pp. 532-533.
51. Христов, И. Битие и съществуване в дискусията за метода между св.Григорий Палама и Варлаам [Christov, I. Being and Existence in the Discussion on the Method between St. Gregory Palamas and Barlaam] In *Хуманизъм. Култура. Религия* (Sofia: «Лик», 1997), pp. 37-48; translated into Russian by Ilya Bey as "Бытие и существование в дискуссии св. Григория Паламы и Варлаама о методе", In Г. Каприев, И. Г. Бей и др. (сост.), *Современная болгарская патрология. Сборник статей*, (Национальные патрологии), Киев: Издательский отдел Украинской Православной Церкви, 2016, pp. 125-138.

Abbreviation

PG – Patrologiae cursus completus. Series graeca. Accurante J.-P. Migne.

Acknowledgment

This research was carried out with a financial support of the Russian Science Foundation, project 16-18-10202, "History of the Logical and Philosophical Ideas in Byzantine Philosophy and Theology."

Notes

1. Octoechos, theotokion dogmatikon, tone 7 (ἦχος βαρύς), inc. Μήτηρ μὲν ἐγνώσθης. Cf., for bibliography, [12, p. 425], [47, p. 244]. Cp. the complete text in English translation by Fr. Lawrence (Campbell) of Jordanville (later monk John): "Thou wast known as a Mother beyond nature, O Theotokos; Yet thou didst remain a Virgin beyond reason and understanding; no tongue can expound the marvel of thy child-bearing; for while thy conceiving, O Pure One, was wondrous, the manner of thy child bearing cannot be comprehended, for wherever God wills the order of nature is overthrown. Therefore as we all acknowledge thee to be the Theotokos we implore thee insistently: Intercede that our souls may be saved." The traditional attribution to John of Damascus is not certain but, at least, corroborated with the manuscript tradition [50].
2. For the overwhelming "Aristotelian" rationalism of John Philoponos (*ca* 490-*ca* 570) that resulted into his so-called "Tritheism", [46], [11]. Cf., for a larger historical context, [35, *passim*].
3. For Barlaam the Calabrian's (*ca* 1290-1348) logical scepticism in theology – an attitude diametrically opposite to that of John Philoponos – [41], [42], [15]. From a logical point of view, the most detailed explanation of the difference between the approaches of Barlaam and Gregory Palamas is provided by Ivan Christov (the only scholar who approached the sources having a logical training in background) [51].

4. Whether these connectives have the same meaning in all possible logics, is a controversial matter and the core of the modern discussion on the logical pluralism, namely, pluralism about the very notion of logical consequence; cf., for a pluralist viewpoint [7] and for a monist viewpoint [39, pp. 194-209]. Be this as it may, for the Byzantine thinkers, a fundamental unity of *the* logic on the level of connectives – but not on the level of the Aristotelian so-called "laws" – seems to me certain.
5. Cf., as an up-to-dated introduction to the field [43]. As a short introduction [38].
6. Cf., for a review of different theologies relevant to Byzantium [35].
7. See for details [28], [2], [3].
8. As a textbook view of the pre-Iconoclastic epoch, I would quote the definition of the anonymous florilegium *Doctrina Patrum de incarnatione Dei Verbi* (*ca* 700): οὐδὲν γὰρ ἕτερόν ἐστιν ἡ ὑπόστασις κατὰ τοὺς θεοφόρους πατέρας ἢ οὐσία μετὰ τῶν ἰδιωμάτων "thus, the hypostasis is, according to the God-bearing Fathers, nothing than the essence with (its) features" [10, p. 72.1-2].
9. Cf. [35], [29], [18], [20], [21].
10. See Dirk Krausmüller's series of three articles on Nicetas Stethatos [19], [17], [16].
11. For the late ninth century, see below, section 8, for Photius's attitude.
12. For a detailed account of his Christological ideas, see [35], [29].
13. Nicephorus of Constantinople, *Antirrheticus* I, 22; [36, col. 252 B]: οὐδεὶς γὰρ τῶν νοῦν ἐχόντων ἀποφανεῖται, οὔτε τὸν Λόγον παθήματα φέρειν, οὔτε τῆς σαρκὸς τὰ θαύματα ὑπολήψεται.
14. Ὁ μονογενὴς Υἱὸς καὶ Λόγος τοῦ Θεοῦ… σταυρωθείς τε, Χριστὲ ὁ Θεός…
15. *Antirrheticus* III, 18-19; [44, cols. 397 D-400 A].
16. Cf. "First then this at least is obviously true, that the word 'be' or 'not be' has a definite meaning, so that not everything will be so and so" (*Metaphysics* IV.4) [5, p. 1589].
17. *Nouveaux essais sur l'entendement humain* IV, 2 [25, p. 361].
18. See [8].
19. Cf. [37], [49].
20. On the actistism [35], [28].
21. See especially [22].
22. Cf. [45].
23. For Damian and his triadology [11], [35]. The most complete bibliography in [27].
24. [14, pp. 44-45].
25. Basil of Caesarea, *Adversus Eunomium*, 2 [6, col. 640.27-30].
26. A number of Bohr's works on this principle are addressing an audience with philosophical interests but without any special training in physics. See most of them in [40].
27. Cf. [1], [9]. To this approach focused on contradiction, add the treatment of identity [13]. Strictly speaking, these logics are inconsistent in a different way than that we are dealing with now; cf., for a detailed review [34].
28. As the best analysis of the doctrine in question [23].
29. Cf., for the problems related to the order in the Trinity or the lack thereof [31] and [26].
30. Cf., for in what sense "three" in the Holy Trinity could be called "number" and similar problems [26].
31. See [33].
32. S. an analysis within the historical context in [4].
33. [48, p. 15.27-28]: λέγομεν ὅτι ἀνελάβετο μὲν τὴν ἀνθρωπείαν φύσιν, ἐξ ἑαυτοῦ δὲ ὁ λόγος παρέσχε τὰ ἰδιώματα.
34. The dogmatic views of Photius are studied surprisingly little, especially in their central topic, Christology. Therefore, below, we will interpret one short text by Photius in a very preliminary manner.
35. "For if Christ had assumed a general man, this would mean that He did not become man in reality or in sensual [perception] but only in thought and imagination (for such is the existence of general things). And in this case He would not have been circumscribed in space according to human nature, for general things are not circumscribed in space" (tr. by Baranov [33, p. 372]); εἰ μὲν γὰρ τὸν καθόλου ἄνθρωπον ἀνελάβετο ὁ Χριστὸς, συμβαίνει αὐτὸν μὴ καθ' ὕπαρξιν μηδ' ἐν αἰσθήσει γενέσθαι ἄνθρωπος, ἀλλ' ἐπινοίᾳ μόνῃ καὶ φαντασίᾳ, αὕτη γὰρ ἡ τοῦ καθόλου ὕπαρξις (lines 3-7; [48, pp. 14-15]).
36. [24, pp. 166-172/130-134 (txt/tr.)]. The relevant chapter II, 18 is entitled "Investigatio confutationis clare significans hanc assertionem: 'Christus est in duabus substantiis secundum commune substantiae [οὐσίας-*B.L.*] significationem (ܡܫܡܥܝܬ ܡܠܬܐ ܕܟܝܢܐ)' ad stultissimam ducere blasphemiam, *scilicet* ad id, quod sancta Trinitas toti humanitatis generi incarnata censeatur" [24, pp. 166/130]. For further literature, see above, note 9.
37. [48, p. 15.8-11]: ἔτι δὲ καὶ εἰ τὸν καθόλου ἄνθρωπον ἀνελάβετο ὁ Χριστός, τὸ δὲ καθόλου τοῦτό ἐστιν, τὸ ἐν πολλαῖς ὑποστάσεσι θεωρούμενον, ἔσται ὁ Χριστός, ἐπεὶ τὸν καθόλου ἀνείληφεν ἄνθρωπον, ὑποστάσεις πολλαί, μᾶλλον δὲ ἄπειροι ("And also: if Christ assumed the man in general, then, given that the general is seen in many hypostases, Christ would be, since he assumed the man in general, many hypostases or, more exactly, infinite [number thereof]").

Studia Humana
Volume 8:1 (2019), pp. 114—116
DOI: 10.2478/sh-2019-0009

Reflections on the Inaugural Conference of the International Orthodox Theological Association (IOTA)

Rico Vitz is the Professor and Chair of the Department of Philosophy at Azusa Pacific University, and serves as the Executive Vice President-Treasurer of the Hume Society. He is the author of *Reforming the Art of Living: Nature Virtue and Religion in Descartes's Epistemology* (Springer), and the editor of *The Ethics of Belief* (Oxford) and of *Turning East: Contemporary Philosophers and the Ancient Christian Faith* (St Vladimir's Seminary Press). He is a member of St. Peter the Apostle Antiochian Orthodox Christian Church in Pomona, California, U.S.A.

Tudor Petzu: Dear Dr. Rico Vitz, as far as I understand, this is for the first time that you are visiting Romania. And I would be tempted to mention that such a visit is a truly blessed one considering that you came to Iaşi, the capital of the ex-Kingdom of Romania and a symbol of Romanian spirituality. That's why I am asking you to tell me how was your first contact with Romania and how meaningful was the spiritual experience of being in Iaşi.

Rico Vitz: Greetings, Tudor. It's nice to speak with you again.

That's right. I had been to a number of European countries, but I had never been to Romania. So, I was looking forward not only to attending the inaugural conference of the International Orthodox Theological Association (IOTA) but also to seeing the country. My experience of Iaşi was wonderful. It is a lovely city with a rich history, both political and religious. And since the conference took place shortly after the Feast of the Nativity, the city was particularly charming in its Christmas decor.

Honestly, the spiritual experience of being in Iaşi and visiting some of the local monasteries was the most meaningful aspect of my time in Romania. On the one hand, this was due to the conference itself since it was a remarkable blessing to have the opportunity to pray and to dialogue not only with Orthodox scholars from around the world but also with the non-Orthodox "ecumenical observers" in attendance. On the other hand, this was due to the rich presence of Orthodox Christianity in the city and its surrounding areas. It seemed as if there were an Orthodox treasure at every turn: e.g., a cathedral, a church, a monument, a museum, a bookstore, a roadside shrine, and so forth. When I came home, I explained my experience to my family and friends this way: the presence of Orthodoxy in Romania is like the presence of Starbucks Coffee shops in the U. S. In Romania, Orthodoxy is everywhere!

Tudor Petzu: It's well-known that you are an American philosopher converted to Orthodoxy so I would like you to explain to me what's the main treasure of Romanian Orthodoxy that you have found. In other words, how would you characterise your experience as American convert to Orthodoxy on the soil of a traditionally Orthodox country?

Rico Vitz: What I found most remarkable was the way that everything in Romania seemed to be, so to speak, "fully saturated" with Orthodox Christianity. As I said, at every turn I encountered a richness and depth of faith both in things, like churches, and in people, not only among monastics and clergy but also among the laity. And I encountered it not only in the more deeply devout and pious, but also in those whose families are still struggling to recover the wealth of personal faith that the communists tried so mightily to steal.

Tudor Petzu: We know very well what a Romanian can learn from America (and there are truly a lot of things, especially if we should talk about freedom and democracy), but what can an American learn from Romania, especially from a spiritual point of view?

Rico Vitz: In all honesty, I am particularly in awe of the Romanian people for their ability to have maintained any semblance of faith despite the immense persecution they faced over the past century. I am also moved by the kindness of nearly every Romanian person I have met both in the U. S. and in Romania. It's really a rather remarkable pair of traits to find in people: that is, both to be so firmly resilient and to be so kind. On these points, we Christians in the U. S. have much to learn from our Romanian brothers and sisters, especially in these acrimonious times.

Tudor Petzu: If you should talk to your American friends about the Relics of Saint Paraskeva, what would you tell them?

Rico Vitz: I've tried to convey both the aesthetic sense of what it was like to be in that holy space and the meaningfulness of the experience. Conveying some semblance of the aesthetic sense to my friends and family in California has been reasonably easy. A number of them have been blessed with the opportunity to venerate the incorrupt relics of St. John Maximovitch at Holy Virgin Cathedral in San Francisco. They have some sense of what it is like to enter an awe-inspiring nave and to approach a holy shrine. For those who have not, I've done my best to describe the beauty and sanctity of the Metropolitan Cathedral in Iaşi.

To understand the meaningfulness of the experience, one has to have an understanding both of the life of St. Paraskeva and of the significance of the communion of saints. Reading and re-telling the life of St. Paraskeva has been useful for helping others (and for helping me!) to appreciate this amazing woman of God. As for understanding the communion of saints, most of my traditional Christian family and friends "get it," but I suspect some of my Protestant friends do not. God willing, in time, those who don't will come to understand and appreciate more fully the significance and power of the "great cloud of witnesses" whose earthly lives are a witness to us and whose heavenly intercession is a blessing for us.

Tudor Petzu: Do you think that this international conference organised by IOTA in Iaşi was an opportunity for Orthodox American philosophy to become more well-known in a traditionally Orthodox country, such as Romania?
Rico Vitz: Yes, I believe the inaugural conference of the International Orthodox Theological Association (IOTA) did present an opportunity for the work of English-speaking philosophers, especially those of

us from the U. S. and the U.K., to become more well known in central and eastern European countries and elsewhere where Orthodox Christianity is the majority religion. I also believe that it presented an opportunity for us, English-speaking philosophers, to become better acquainted with the philosophical and theological work that is being done by our academic colleagues and spiritual brethren of those countries.

My hope is that the conference will lead to greater interaction and collaboration among us all, and that this work will make meaningful contributions in our efforts to work out our salvation, by loving God and loving our neighbors. That is my hope. Will it be realized? We'll see, as God wills it.