History of Medicine, 2019, 6(4): 191–199
DOI: 10.17720/2409-5834.v6.4.2019.01a

Artificial morality for artificial intelligence

Nikolay N. Krylov[1], Yevgeniya L. Panova[1], Aftandil V. Alekberzade[1]

[1] *FSAEI HE I.M. Sechenov First MSMU MOH Russia (Sechenov University)*
2 Bolshaya Pirogovskaya St., building 4, Moscow 119991, Russia

Corresponding author: Nikolay N. Krylov (nnkrylov01@yandex.ru)

Received: 15 December 2019 **Accepted:** 30 December 2019

Citation: Krylov NN, Panova YL, Alekberzade AV (2019) Artificial morality for artificial intelligence. History of Medicine 6(4): 191–199. https://doi.org/10.17720/2409-5834.v6.4.2019.01a

Abstract

To unify the solution to the problems faced by the creators of algorithms for artificial intelligence (AI) for making moral decisions, both multifarious variants of speculative experiments and the results of studying the consequences of real events or generally accepted actions and stereotyped decisions were proposed. As a general rule, these were models of various critical situations requiring immediate solutions and designed to test the range of problems arising in the course of practical use of artificial intelligence in the field of administration and security. Various moral dilemmas, both artificially created and based on real events, were proposed as models for the decision-making algorithm. Decision-making requires defining the boundaries of the legitimacy of decisions made by AI. The authors analyse the logic of the choice between life and death in the 8th declamation of pseudo-Quintilian, as well as in the Survival Lottery (an experiment with organs for transplantation), the Terrorist Ultimatum, the trolley problem, and in the Moral Machine problem. Life forces us to constantly make choices to solve a wide range of everyday tasks, such as clinical experiments of physicians, medical triage of the wounded on the battlefield, treatment of patients in a state of prolonged coma and with orphan (rare) diseases, and other problems upon which the fate and lives of people depend. The authors are convinced that, at present, there is no universal morality that could serve as the basis for the creation of AI, including that for driving vehicles. When creating a universal morality for AI, one should consider the answer to the main question: do lives of all people have the same value?

Keywords

history of medicine, universal morality, artificial intelligence, 'trolley problem', 'moral machine', real moral dilemmas

In the 1950s, expert systems describing the algorithm of actions for choosing a solution depending upon specific conditions began to appear in Russia and the rest of the world. In 1961, a laboratory of medical cybernetics was created at the A.V. Vishnevsky Institute of Surgery, and the research of the issues of diagnostics and prognosis of diseases (and, later, the research into remote diagnostics using teletype communication) with the help of computers began. Gradually, machine learning came to replace expert systems, which now allows us to talk about the emergence of artificial intelligence (AI).

Modern man is guided by a variety of moral grounds. Apart from public opinion and education, such grounds can also include one's own moral principles based upon philosophical ideas and religious beliefs, as well as the requirements of the professional environment and legal restrictions existing within a certain time frame. Such a multitude of problems inevitably leads to conflicts and contradictions, which is also an important ethical problem in its own right.

The development of AI is inextricably associated with the algorithmisation of the actions of an autonomous system in conditions when a specific decision is made by the said system instead of a person (an operator or a driver), or when this decision can directly or indirectly affects lives (or health) of people. In this case, it becomes necessary to create some uniform rules for the functioning of AI that satisfy the beliefs of all (or most) people in making moral decisions.

To unify the solution to the problems faced by the creators of algorithms for artificial intelligence (AI) for making moral decisions, both multifarious variants of speculative experiments and the results of studying the consequences of real events or generally accepted actions and stereotyped decisions were proposed. As a general rule, these were models of various critical situations requiring immediate solutions and designed to test the range of problems arising in the course of practical use of artificial intelligence in the field of administration and security. Various moral dilemmas, both artificially created and based on real events, were proposed as models for the decision-making algorithm. Another important aspect is the determination of the boundaries of the legitimacy of decisions made by AI,[1] i.e. restrictions necessary to prevent errors resulting from the functioning of AI or failures in its work.

Speculative experiment

We receive our life not of our own free will and without our desire, but we lose it, as a general rule, imperceptibly to ourselves, like a bygone day at sunset. The conscious choice between Life and Death is simple and inherent only in people. Life or Death – both an alternative and a dilemma – is clear to everyone. Everyone has to choose between them (either constantly, due to professional responsibilities, or situationally, at least once). The most sophisticated options for making such a decision remain in people's memory as traces of the formation of morality.

The logic of the choice between life and death in reasoning abstracted from life, perhaps, first appears in the 8th declamation of pseudo-Quintilian (2nd century). It tells a story of a family in which two twin brothers fall ill. The parents go to a doctor, and it becomes clear that the children suffer from the same disease, the essence of which is unclear to the physician, so both brothers are doomed to die. To at least save one of the twins, the doctor suggests that the other one be vivisected to diagnose the disease. The father agrees to this condition. The doctor dissects one of the twins and examines his organs, whilst the ailing one, still being conscious, cheers the doctor. One of the children dies, but the other is cured (Ferngren 2017). The story is uncomplicated but loaded with dramatic details to highlight the problem of a double choice of the correct solution: the need to agree to kill one of the children, and also to choose which of the two brothers will live. The obvious contradiction of one dilemma with human nature is emphasised by the impossibility of a rational solution to the other.

Later, the dichotomy "life in exchange for inevitable death" was used to study the foundations of morality. The simplest versions of a speculative experiment are presented in the discussions of the Survival Lottery, as well as the Terrorist Ultimatum. The idea of the philosopher John Harris (Harris 1975) revolves around the initial adoption of a number of assumptions: organ transplantation is highly effective, organs from one donor can be used to save the lives of at least two recipients (currently from 4 and up to 8), death from natural causes is equivalent to death for donation (i.e. killing the donor). According to the author's assumption, every living person has some kind of identification number. When it becomes necessary to save the terminally ill, a computer runs an unbiased lottery amongst healthy people to randomly select a donor based upon their number. The chosen one *a priori* agrees to donate their organs, including unpaired ones (heart, liver), i.e. sacrifice their life in order to extend the lives of other people (two or more). An obvious quantitative gain may testify in favour of the fictional viability of such conditions: two (or more) lives saved compared to one lost (completely healthy and presumably longer), as a result of a random choice. Is it possible to accept, if only mentally, such a condition and agree with its hypothetical result?

The conditions of the issues of A. Campbell and P. Foot are different to the one above in form, but close in essence. The test subject must decide on the appropriacy of agreeing with the terms of the ultimatum of a tyrant (invader, terrorist): to either kill two combatants himself in order to save 80 hostages from amongst civilians (Campbell 2013, p. 21-22), or torture only one in order to save 5 innocent people from the terrorist (Foot 1967). The conditions of such moral dilemmas are formally plausible, but the conditions of the test are far from the usual realities of everyday life. And, most importantly, the authors of the question "suggest" the obvious advantages of the utilitarian approach (saving more is better), thus reducing the test subject's intelligence to the level of an adding machine and depriving him of the opportunity to leave the game and save face.

It is obvious that solving such problems has not become an AI training programme and does not claim the universality of the pseudo-morality embodied in them. Such dilemmas are too crude a tool for learning to make infallible moral judgements. This fact seems to be the reason why these and similar moral dilemmas have not become widespread. But an attempt to complicate them has led to the emergence of fundamentally new problems – the trolley problem (the trajectory of the trolley/tram) and the Moral Machine.

Philippa Foot's ethical puzzle (Foot 1967), known today as the trolley problem, revolutionised the study of morality in a way, giving impetus to a large amount of research in ethics, philosophy, AI and clinical medicine, revolving around the use of this striking, memorable dilemma (Thomson 1985).

According to the conditions of the problem, a heavy runaway trolley is barrelling down the railway tracks, and it can continue to move in only one of two sets of tracks. Ahead of it, on the main track the trolley is already set for, there are five people tied up and unable to move, and there is only one person on the

[1] See: http://publication.pravo.gov.ru/Document/View/0001201910110003. Access: 07.12.2019.

side track. The test subject can either be in control of the steering wheel of the trolley or the lever that can switch the trolley between the tracks. As a result, the person controlling either the trolley or the switch (the testee) must decide which path the trolley will take, i.e. whether it will kill five people and save one, or kill one person and save five. This task has many different options, each of which raises ever new ethical questions and renews ongoing debates (Tversky and Shafir 1992; Greene et al. 2009; Kortenkamp and Moore 2014). The new pre-emerging circumstances – the possibility to stop an uncontrollable trolley by pushing a very fat man onto the rails – are the ones being discussed most frequently (Thomson 1976). According to the terms of the problem, his weight would be enough to save everyone lying on the tracks, but in this case, he must be sacrificed (the person in question is sick, flawed and defective). Additional options for studying the reactions of the tested person are to hold the trolley control levers in their hands in order to direct it to a living obstacle (a fat man on the tracks) or to throw that unfortunate person on the track using a special device.

These classic versions of the test (Andrade 2019) can be made more complex with the addition of certain circumstances. So, for example, switching the arrow takes the life of not just an abstract, depersonalised entity, but a perfectly specific person, well-known to the test subject (their relative, loved one, or an acquaintance). In another case, the circumstances are "mitigated" by the fact that the one lying on the tracks, as it transpires, is an outstanding writer, inventor, scientist or physician. Will this circumstance simplify the choice, make it more obvious? Most likely not. But as soon as it is announced that the "fat man" standing next to the railway tracks is the criminal who devised such a terrible experiment, the moral assessment of the situation changes again and it becomes much easier to sacrifice the villain without a trial.

In another study, test subjects willingly sacrificed themselves, "throwing themselves onto the tracks" as an obstacle to the trolley in those cases when it was necessary to prevent the death of a group of people they identified with (Latin Americans) (Swann et al. 2010).

There is an obvious way to increase the list of options for formulating this problem to infinity. Five people are tied to the tracks and will die if the lever is not switched. However, if the test subject does this, then the trolley will go along the second track, where the situation will repeat itself, but not with five, but ten people tied to the tracks, and yet another person will have to make the same choice. The number of dichotomies on the tracks is infinite, the number of possible victims will increase multiples times each time. So, in the end, the entire planet can become depopulated because of the desire to just once refuse the wrong choice!

Comparing the principles of formulating the dilemmas of J. Harris and P. Foot, one can find a commonality of formal features. However, given the seemingly obvious similarity of the tasks' scenario ("how many people to kill and how many to spare") – the possibility of obtaining organs for transplantation from a healthy living person to save several patients, and choosing the path for the trolley (save five people by sacrificing one) – the decisions made by most of the tested are directly opposite. This gives rise to the question of why, in one case, the postulate "the greatest good for the greatest number" is taken into account, whilst in the other one the argument "the rights of one innocent person are of far greater importance than the achievement of health for several patients" (Albin 2005). The reason for this is deemed to be the existence of a social contract (Rosenberg 1992).

So, an attempt to resolve such dilemmas exposes the problem of what is more important to us, the final result or how it is achieved. The shortcomings of the tool for solving moral problems – internal inconsistency, sketchiness and simplification – are also obvious. However, it is precisely such dilemmas, first and foremost the trolley problem, that serve as the basis for honing the algorithms for controlling a self-driving vehicle (Epting 2016; Faulhaber et al. 2019).

However, R. Davnall is convinced that if a situation similar to the trolley problem arises on the road, provided that the necessary information is available to the car computer (including the magnitude of the braking dynamics and tyre traction), the decision should not be associated with measuring the trajectory of movement and running people over. For a car, as experience shows, it is always the least risky to brake in a straight line and not turn to the side (as the conditions of the dilemma recommend) (Davnall 2019).

The presented dilemmas, in our opinion, are a simplified reflection of reality, and not a guide to action. Besides, in a critical situation, the driver's actions are often intuitive, reflexive, and not rational. The driver develops and hones the necessary reflexes – skills, automated actions, including actions in a critical situation – at the training ground. Traffic codes teach us to avoid getting into an accident. Therefore, the AI of a self-driving vehicle should contain exactly that information and such algorithms, and not a cold count of the number of the likely injured (or killed).

It is obvious to us that the ethical scenario of the trolley problem cannot be a key criterion for making moral decisions in difficult ethical situations in medicine, since the comprehension of all the options for its movement and methods of stopping is not associated with the choice of options for saving people, but rather turns into a search for a sophisticated, comforting way of creating conditions for murder; they impose upon the Everyman a solution associated with death, postulating its inevitability; and murder in modern civilised society is taboo itself.

The extrapolation of such decisions in medicine is unacceptable, since "a doctor can neither be a tormentor, nor an executioner, nor an executioner's servant" (Paracelsus). But such dilemmas serve as a certain starting point of initial intuitive guidance for finding a way out of a difficult professional situation, for example, such as abortion and euthanasia (Andrade 2019), treatment of Ebola (Lally 2015), cardiopulmonary resuscitation in emergency surgery (Manthous 2014), transplantation of internal organs (Kolber 2009), cognitive neurology (Cushman and Greene 2012) and research of the problem of hyperdiagnosis in clinical medicine (Carter 2017).

Nevertheless, it is believed that the declared morality cannot be applied to justify the occurrence of side effects and their consequences during mass vaccination, and the general willingness of test subjects to sacrifice other people's lives in the context of a "fair trolley" may be an artefact of an unrealistic setting of the task (Dahl and Oftedal 2019). Therefore, special caution is needed in using artificially created dilemmas as a key testbed for identifying the essential foundations of moral judgement (Rai and Holyoak 2010). These shortcomings, as well as the small number of the tested, insufficient to obtain reliable conclusions, have served as the basis for the creation of another version of the speculative experiment – the Moral Machine.

The Moral Machine (MM) is the largest research test we know of, designed to examine the moral dilemmas that self-driving vehicles may face. Within a year and a half, this online experimental platform managed to get the results of answers to questions about more than 39 million moral decisions of people from 233 countries and territories in ten languages of the world (Awad et al. 2018). The second similar work collected more than 12,000 moral decisions from several thousand people from the USA and Denmark (Frank et al. 2019), the third – more than 18 million answers from more than 1.3 million tested (Noothigattu et al. 2017).

MM employs a decision-making methodology that is conceptually related to the trolley problem scenario. The test subject can direct a car with passengers either to an obstacle on the road (a crash test that, according to the conditions of the problem, always results in the death of everyone in the car), or to pedestrians, also involving the obligatory death of everyone in the way, but saving the passengers. There is no third option! Instead of a human, the solution to this dilemma can be redirected to the AI controlling the self-driving vehicle that is pre-programmed to make such decisions.

Unlike the trolley problem, MM can offer significantly more options for situations on the road and in the car – up to 56. In each of them, one of the two alternative outcomes (death of pedestrians or death of passengers) was analysed in the form of a multifactor design of numerous dilemmas with various situational factors: the number of passengers (1, 2 or 4), the number of pedestrians (1, 2 or 4) and their age (old and young), the presence of a child amongst passengers or pedestrians, as well as violation of traffic rules by one or more pedestrians. In addition to that, supplementary characters and details could also appear in the tasks: pets (a dog), gender of potential victims (men and women), their social status (high and low), profession (athlete, doctor), physical status (pregnancy), social deprivation (criminals, homeless people).

According to the authors of the project, they've managed to identify some "global preferences", such as the preferential preservation of the lives of people, not animals, or a greater number of people; and the choice in favour of the young (Awad et al. 2018). On the other hand, they haven't established a pronounced influence of the individual characteristics of the subjects (age, education, gender, income level, political and religious views) on the test results, and the revealed insignificant variations are more of a theoretical rather than practical significance.

At the same time, they've discovered the existence of three "moral clusters" of countries united into separate groups based upon geographic and cultural parameters. The first ("western") cluster includes North America and many European countries of the Protestant, Catholic and Orthodox Christian persuasion. The second ("eastern") one includes the countries of the Far East of the Confucian cultural group (Japan and Taiwan) and Islamic countries (Indonesia, Pakistan and Saudi Arabia). The third cluster ("southern") united the countries of Central and South America. The differences in culture and traditions of these regions, the level of social development and ideas of egalitarianism, as well as in the assessment of the usefulness and value of life between these conditional associations of countries, were manifested, amongst other things, in the preferences of the choice of their participants regarding the worship of the elderly and people who have a higher social status, the need to save the lives of passengers (not pedestrians), women (not men) and pets.

This delivers the message that the task of constructing machine morality (AI decision algorithm) should be based upon the moral decisions of people, and is essential to take into account a wide range of their biases, which have a different basis.

D-A. Frank and co-authors used contextual factors that were similar and close in meaning, but variation in research subjects (passenger, pedestrian, bystander) and dilemmas (28 in total) increased the number of moral decision options in some problems to 84 for each participant (Frank et al. 2019).

This design of the study made it possible to confirm the hypothesis about the absence of both the unchanging *status quo* and moral status of the individual, and the bias in the decisions made during the experiment (both for an individual person and in a relatively culturally homogenous selected group of people). Given the formulation of an abstract problem, testees usually express an intuitive moral preference to sacrifice the self-driving vehicle and its passengers, rather than harm innocent pedestrians. They also maximise the usefulness of the lives saved by sacrificing a group with a smaller number of people, and tend to spare the lives of children, and they are significantly more likely to sacrifice pedestrians who violate the red light stop rule, demonstrating a utilitarian approach. However, as soon as the formulation of the assignment changes to have a personal connotation and the situation revolves around the test subject themselves or their relative, who is in the car or crossing the road, the moral decision about who should be sacrificed radically changes exclusively in favour of the member of "their group", regardless of the environment and circumstances, now professing the deontological doctrine.

Thus, the development of a universal moral code for AI in general and for self-driving vehicles in particular allows for the existence of global moral preferences, which are based upon easily identifiable culturological axioms and generally accepted rules of behaviour.

Real (precedent) examples of moral choice

Obviously, the study of the Moral Machine problem and, furthermore, the trolley problem, is not realistic as it lies outside the boundaries of everyday life. It is impossible to imagine how, in a real situation on the road, a vehicle (a tram or a car), at the behest of the driver, would be faced with the actual choice between two different types of people. Most likely, a participant in such an experiment forgives themselves unpleasant ambiguous moral decisions because they view themselves not so much "inside" the scenario, as a direct participant in the tragedy, as "above it" – as a player in the attached circumstances. A person who has been behind the wheel of a car at least once obviously realises that they must head towards the massive obstacle (if there is no possibility of lateral braking with the separation barrier), relying upon seat belts and airbags to save the lives of passengers so as not to injure pedestrians, whoever they are. The statement that everyone in the car is always better protected than pedestrians does not require proof!

However, life, in contrast to a vulgar game, constantly invites us to choose one task from a wide range of everyday tasks. Easy decisions generally do not remain in memory and do not touch upon the fundamental issues of morality; complex decisions are kept in memory for a long time and can become an example of moral choice in similar conditions.

The history of mankind is rife with descriptions of the altruistic actions of soldiers who sacrificed their lives in order to save the lives of their fellow soldiers ("There is no greater love than to lay down one's life for one's friends", John 15:13). It is impossible to even think that in critical moments a person is capable of cold-blooded calculation, i.e. to decide whether one death (their own) is comparable to the death of five (or ten) of their enemies, and weigh *pro et contra*.

At the same time, for example, the actions of military pilots in peacetime, when they diverted a falling plane from residential buildings and saved people at the cost of their lives, fully fits into the idea of fulfilling military duty, and not the requirements of the deontological maxim.

The numerous experiments of physicians, when, in order to obtain the result necessary for further research, they artificially infected themselves with various diseases or placed themselves in deadly experimental conditions (for example, asphyxiation, starvation, fresh water deprivation), are all examples of self-sacrifice in the history of medicine.

Which decision is easier to make? To sacrifice the "life" of a non-existent person invented for a game, for a speculative experiment, or one's own, for the good of other people, or to take on the role of God and decide whether a particular person will live on or not?

It is likely that Louis Pasteur, who wasn't a physician, had a very difficult time coping with an insoluble dilemma when he first tested the rabies vaccine in 1885. He gave his permission to administer twelve doses of the drug to nine-year-old Josef Meister, but he was still confronted with two variables that determined the fine line between his personal success and failure: the drug was effective in experimental animals, but it was unclear how it would react with an infected person; despite the anamnesis and the objective signs of rabid dog bites, it was unclear whether the boy was infected or not. Possible combinations of these unknowns led to four different scenarios for the first clinical trials of the new drug. In order to be convinced that the patient is actually ill, Pasteur decided to make the 13th injection, but with a wild, not attenuated culture, thus definitely infecting the patient (in case a rabid animal's bite is present, but no infection has occurred). If the vaccine were ineffective, the patient would have probably died from an incurable disease. Pasteur was aware of this, but he won by betting authority – but not his life. The obvious immorality of this act cannot be overshadowed by noble thoughts, although, from a utilitarian point of view, a risky clinical experiment on one person made it possible to save the lives of many thousands of people in the future. So is the win obvious?!

Academician A.A. Smorodintsev experimented on his own granddaughters at the stage of vaccine testing, injecting one with an attenuated strain of the poliomyelitis virus and the other with measles. This was done with the knowledge of crying parents and, formally, could meet the requirements of the WMA Declaration of Helsinki. Unlike J. Meister's mother, they understood all the risks and severity of the likely consequences, but in the USSR, it was customary to be proud of this act of the immunologist. The researchers at the L. Pasteur Institute (now the Saint Petersburg Pasteur Institute of Epidemiology and Microbiology) only followed the tactics of the great Frenchman, despite its obvious immorality. However, it is easy to preach a specific moral norm as a generally accepted imperative, but it is much more difficult to follow it, including in medicine.

More often than not, throughout the history of mankind, a brave doctor (as opposed to a soldier or a pilot) would put other people's, and not own, health and life, at risk. But making a decision in favour of one or the other patient (the choice of which patient can still be saved, and which should be deemed hopeless) is not easier, if not more difficult, than the right to dispose of one's own life.

A real working model of decision-making in military and disaster medicine is mass triage of the wounded, first developed in the 19th century by Dominique-Jean Larrey, N.I. Pirogov and Johann Friedrich August von Esmarch. In this case, instead of the question of which of the innocent people it is better to kill (as is the case of the trolley problem or MM), there is a question of who can and should be saved. Despite the seeming equivalence, the situation changes dramatically, since in this case the doctor builds a future strategy towards reducing the total amount of suffering. Most importantly, the personal ethical standards of a physician should not diverge from the ethical requirements of society in conditions of extreme emotional overload, since there is a shortage of medical resources and/or the impossibility to provide timely, full-fledged medical care to a large number of differently affected people in a short period of time. At the same time, in one of three or five groups of victims (depending upon the triage option), it is necessary to select the category of "the hopeless", whose life – according to the prognosis at the moment – cannot be saved under the current circumstances due to injuries incompatible with life. However, the doctor should show compassion for such patients, respect for their human dignity and their lives (place them separately from others and prescribe analgesics and sedatives). In the context of making a life-determining decision, the doctor, most likely, latently believes in the inconclusiveness of their own verdict, since over time the objective condition of the wounded changes and, carrying out mandatory repeated monitoring, adjustments should be made to the initial assessments, which inspires hope and removes the harshness of the final conclusion.

The medical decision-making on the selection of patients who will undergo haemodialysis from the group of patients suffering from terminal chronic renal failure should be considered close in form to triage. In the recent past (mid-20th century), there were significantly more candidates for treatment with an artificial kidney device than the capabilities of the available devices. Nowadays, a specialist also selects patients for the treatment of rare life-threatening chronic progressive diseases. Of the 5,000 – 7,000 rare ailments existing in the world, only 219 of them are placed in the list of the Ministry of Health of the Russian Federation, and only in the case of just 24 of them patients with "orphan" diseases are fully provided with the necessary drugs and nutritional therapy by the state. Given the obvious limited funding for this budget line, it becomes problematic to prescribe, for example, Zolgensma – a drug for gene therapy for children under two suffering from spinal muscular atrophy with mutations in the SMN1 gene, worth more than US $2 million. In fact, the decision to prescribe a specific expensive treatment for a particular patient with an orphan disease (haemodialysis in the recent past) or to refuse to do so is an example of a classic moral dilemma.

Of great relevance are also the problems of choosing a recipient for transplantation of internal organs from amongst many candidates (no more than 8-10% of those in need can be provided with donor material) and the separation of conjoined twins (the surgeon risks the life of either one or both of them at once).

The extreme demand for donor material is evidenced by the fact that the illicit trafficking in the organ transplant market is estimated at US $600 million: potential recipients do not want to wait for their chance to live, which may never come.

The twin sisters Jodie and Mary were born fused below the waist, with the shared heart, lungs and liver located in Jodie's body. In November 2000, at St Mary's Hospital (Manchester, England), doctors gave only one of them the chance to survive, taking Mary's life and separating her body from Jodie's. The dilemma in this case boiled down to the following: should both girls die, or just one? In 2003, at the N.F. Filatov Children's City Clinical Hospital No. 13 in Moscow, Russian surgeons led by Academician Isakov performed a ten-hour operation to separate eleven-year-old ischiopagus twins Sita and Gita, who were fused at the pelvic area, had a shared bladder and three legs. In this case, the operation was successful and the unfavourable outcome was delayed by more than ten years. In 2015, Sita died of multiple organ failure, but Gita carries on living an active life. Each case of such operations is unique, but choosing the right solution is always more difficult than in the case of MM.

The absence of universal, indisputably right signs of brain death, irreversibility of a coma or a vegetative state is clearly demonstrated by the examples of partial recovery of physical and mental activity in patients after prolonged periods of being in a critical condition. Over the years, relatives of such patients repeatedly discuss with medical professionals the need to disconnect life support and stop caring for them. Not abstract, but quite specific reasoning regarding the subject of what to do next – continue caring or let the patient die – is an example of a topical dilemma. However, Donald Herbert, Terry Wallis and Jan Grzebski, who were in a coma for 10, 17 and almost 19 years respectively, like some other patients, regained their ability to communicate with the world.

Reading about difficult medical choices is easier than making them. One evening, before leaving the clinic, one of the authors of this article went to the intensive care unit in order to examine the patient who had been operated on earlier that day. During the round of the department with the ICU nurse-on-duty, the power went out in the entire building due to an electricity failure after repair works on the upper floors of the building. The author had to instruct the colleague: "Work here, in this room, and don't go anywhere – I'll provide assistance in the other room!" At that moment, there were four patients on ventilators in two intensive care rooms: three in one section of the ward (one after palliative surgery for terminal cancer, and two others after radical surgery – the author's patient and one operated on by a novice surgeon), and one in the other section, operated on by the head of the clinic (where the colleague had remained). Effective ventilation was maintained through an endotracheal tube using an Ambu breathing bag. The nurse-on-duty was involved in organising the restoration of power supply.

Every day, the real clinical practice is ready to pose questions more sophisticated than the speculative discourses of an armchair scientist. How should you manage the two doctors? Where should they provide assistance? In this case, should one doctor move from one room to another, or is there no need to waste precious time on such things? The decision made allowed the first doctor to concentrate on one patient, whilst the second moved between the adjacent beds of three critically ill patients. After 15 minutes of intensive work, when the energy was almost depleted, the electricity came back on. Three patients coped with the emergency situation and survived; the patient operated on by the author died three days after the incident (he developed cerebral ischaemia). Didn't help "his own" patient because he had to help the other patients?! Should have concentrated only on the 'right' patient?! What's the right answer?!

These and similar questions, such unresolved medical dilemmas, leave non-healing wounds in the souls of doctors of all specialities in all countries of the world, as well as the "emotional exhaustion" syndrome (Samokhvalov, Krylov, Vychuzhanin 2017).

As a general rule, historians, politicians and the military are the best at handling utilitarian problems, even without the help of AI. "One death is a tragedy, a million deaths – a statistic".[2] So, for example, they explain the atomic bombings of Hiroshima and Nagasaki on the 6th and 9th of August 1945, by humane considerations, since the use of weapons of mass destruction led to the death of 'only' 150,000 – 200,000 Japanese, whereas the continuation of conventional hostilities for several more months would have led to the death of about 500,000 American soldiers. During the carpet bombing of Tokyo alone, conducted by the United States on the 10th of March 1945, between 80,000 and 100,000 people died. The number of casualties of the Allied bombing attack on Dresden between the 13th and the 15th of February 1945, aimed more to intimidate rather than achieve any military goal, was between 100,000 and 250,000 people. How to assess the ethics of such a choice? Then again, politics is only the quintessence of morality accepted in society.

Difficult political dilemmas are tackled quickly and pragmatically. How to tell heretics from righteous people, Catholics from Cathars (Albigensians)? Abbot Arnaud Amaury in July 1209, during the siege of the fortress of Béziers, uttered the sacramental phrase: "Kill them. For the Lord knows who are His" (*Caedite eos. Novit enim Dominus qui sunt eius*). Thus, countless people were killed in the city on that day[3] (Caesarii Heisterbacensis... 1851, p. 302).

Nowadays, people are destroyed in different parts of the world, every day, without coordination with AI and in accordance with some momentary moral preferences, as if in a computer game, "in the name of the interests of the country" or "for the purpose of preventative defence".

Conclusion

Speculative experiments with MM and the trolley could be compared with Zeno's paradoxes (discourses about motion and sets). They are just as notional and abstract, conceived as exercises for the mind; they do not stand up to scrutiny when being extrapolated into reality. However, Zeno's reasoning is distinguished by paradox and philosophical depth, the elaboration of

[2] "One death is a tragedy. A million deaths is just a statistic". Kurt Tucholsky.

[3] Here is the excerpt: "Cognoscentes ex confessionibus illorum catholicos cum haereticis esse permixtos, dixerunt Abbati: Quid faciemus, domine? Non possumus discernere inter bonos et malos. Timens tam Abbas quam reliqui, ne tantum timore mortis se catholicos simularent, et post ipsorum abcessum iterum ad perfidiam redirent, fertur dixisse: Caedite eos. Novit enim Dominus qui sunt eius. Sicque innumerabiles occisi sunt in civitate illa". – "When they discovered, from the admissions of some of them [of those in the fortress], that there were Catholics mingled with the heretics they said the toeh abbot "Sir, what shall we do, for we cannot distinguish between the faithful and the heretics". The abbot, like the others, was afraid that many, in fear of death, would pretend to be catholics, and after their departure, would return to their heresy, and is said to have replied "Kill them all for the Lord knoweth them that are His (2 Tim. ii. 19) and so countless number in that town were slain". (See: https://sourcebooks.fordham.edu/source/caesarius-heresies.asp#CHAPTER%20XXI).

one of the sides of the endless number of facets of life, and the brakeless MM toy, rushing along the motorway, is just an analogy and an allusion to the lack of alternatives to inevitable death.

Liberalisation of modern society and pluralism actually deprived a significant part of Western thinkers of the idea of *telos*, the purpose of man, at the same time secularising both moral choice and moral guidelines (Delkeskamp-Hayes 2015). The distinction between the sacred and the profane has led to a desacralisation of the concept of a righteous, "good life", established convictions, inviolability of moral foundations, and the erosion of ideas about the immoral. This concept claims to be universal and neutral, just like the MM logic (Iltis 2015).

The authors of a number of works make an assumption about the existence of moral preferences of the majority of people – a kind of "universal public morality" – and assume that they can establish its main theses with the use of the declared methods (Awad et al. 2018; Frank et al. 2019; Noothigattu et al. 2017). It is assumed that through the general recognition of such preferences, humanity will be blessed with the knowledge that gives access if not to the morally optimal, then at least to the morally acceptable choice in a situation of uncertainty. The civil contract established in this way, like some stone tablets, like "the highest truth", must issue a moral license for the chosen preferences. One of the forms of realisation of such a syllogism is equipping the AI of a self-driving vehicle with the beliefs of universal morality.

In addition, the impossibility of a spontaneous stopping of a moving vehicle (car, carriage) under the test condition gives rise to an allusion of an unmanageable and uncontrolled passage of time and the possibility of stopping it only through catastrophe and death. Such a plot in the context of solving urgent globalist problems becomes acceptable, tempting, attractive and quite feasible.

Such reasoning, in our opinion, is erroneous, sine all speculative experiments with a trolley or MM take place in a "legal vacuum", under conditions of complete disregard for the existence of traffic rules, whilst in every legal state there are concepts of "criminal homicide" and "unintentional homicide". Criminal legislation is, although maybe not absolute, but nonetheless an expression of the very social morality that is currently accepted as a comprehensive "social contract". We believe that all the moral and ethical decisions tested by MM are made in the world of some kind of virtual reality, a computer game, and can be transferred to life even outside the traffic setting, if the perceived boundaries of the game are lost. In practice, the creators of MM have already assumed the functions of the court, the jury and the executioner (recall the words of one F.M. Dostoevsky's characters: "Am I a trembling creature, or do I have a right?"). The next step in the chain of such judgements should be the moral permission to run people over (kill!) based upon their skin colour, religious, national, sexual, professional, physiological and medical characteristics and age.

From our point of view, each respondent – the test subject of the MM experiment – should, first of all, answer the organiser's question: "Do lives of all people have the same value to me?". It is the answer to it that allows us to:

1) recognise the design and results of the MM experiment as immoral;

2) argue that, at present, there is no universal morality that could serve as the basis for the creation of AI, including that for driving vehicles;

3) expect that with the advancement of AI and machine learning, the framework of modern morality must be prepared to deal with the problems that may arise in connection with these evolving systems, and the design of these automated systems must be adapted to the perception of these moral principles;

4) assume that the creation of AI will lead to a better understanding of human morality.

So, the awareness of the objective shortcomings of moral decision-making models does not allow the use of modern AI solutions to integrate such autonomous technologies into moral spheres, including medicine, law, military forces and self-driving vehicles (Bonnefon, Shariff, Rahwan 2016; Greene 2016; Bigman and Gray 2018).

Artificial Intelligence still "lacks Heart, Soul and Compassion".[4] At present, it is impossible to use the results of earlier speculative experiments to create an algorithm for the moral actions of a self-learning artificial mind. The phenomena of real life are brighter, more diverse, unpredictable and profound than the primitive scenarios and moral foundations of a virtual computer game. Comprehension of all aspects of life is the main issue and the main goal of the creators of the analogue of the human mind.

[4] See: https://tass.ru/obschestvo/7274547. Access: 07.12.2019.

References

Albin RL (2005) Sham surgery controls are mitigated trolleys. J Med Ethics 31 (3): 149-152. doi:10.1136/jme.2003.006155

Andrade G (2019) Medical ethics and the trolley problem. J Med Ethics Hist Med 12(3): 1-15.

Awad E, Dsouza S, Kim R, Schulz J, Henrich J et al. (2018) The Moral Machine experiment. Nature 563: 59–64. doi:10.1038/s41586-018-0637-6

Bigman YE, Gray K (2018) People are averse to machines making moral decisions. Cognition 181: 21-34. doi:10.1016/j.cognition.2018.08.003

Bonnefon J-F, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352 (6293): 1573-1576. doi:10.1126/science.aaf2654

Campbell AV (2013) Bioethics: the basics. Taylor & Francis Books.

Carter SM (2017) Overdiagnosis, ethics, and trolley problems: why factors other than outcomes matter - an essay by Stacy Carter. BMJ 358: j3872. doi: 10.1136/bmj.j3872

Cushman F, Greene JD (2012) Finding faults: how moral dilemmas illuminate cognitive structure. Soc Neurosci 7 (3): 269-279. doi: 10.1080/17470919.2011.614000

Caesarii Heisterbacensis monachi Ordinis Cisterciensis Dialogus miraculorum. Vol. 1 (1851) Ed. Joseph Strange. Köln; Bonn; Bruxelles: Sumptibus J. M. Heberle (H. Lempertz & comp.).

Delkeskamp-Hayes C (2015) The distant echo of Aristotle in bioethics today – and how to reduce the Noise. History of Medicine 2 (4): 431–441.

Dahl FA, Oftedal G (2019) Trolley Dilemmas Fail to Predict Ethical Judgment in a Hypothetical Vaccination Context. J Empir Res Hum Res Ethics 14 (1): 23-32. doi: 10.1177/1556264618808175

Davnall R (2019) Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics. Sci Eng Ethics. Preprint. Published online: 01 April 2019. https://link.springer.com/article/10.1007/s11948-019-00102-6. doi: 10.1007/s11948-019-00102-6

Epting S (2016) A Different Trolley Problem: The Limits of Environmental Justice and the Promise of Complex Moral Assessments for Transportation Infrastructure. Sci Eng Ethics 22 (6): 1781-1795.

Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S et al. (2019) Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. Sci Eng Ethics 25 (2): 399-418. doi:10.1007/s11948-018-0020-x

Ferngren G (2017) Vivisection Ancient and Modern. History of Medicine 4 (3): 211–221. doi:10.17720/2409-5834.v4.3.2017.02b

Foot Ph (1967) The Problem of Abortion and the Doctrine of the Double Effect. Oxford Review 5: 5–15.

Frank D-A, Chrysochou P, Mitkidis P, Ariely D (2019) Human decision-making biases in the moral dilemmas of autonomous vehicles. Scientific Reports 9. Published online: 11 September 2019. https://doi.org/10.1038/s41598-019-49411-7.

Greene JD, Cushman FA, Stewart LE, Lowenberg K, Nystrom LE et al. (2009) Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition 111 (3): 364–371.

Greene JD (2016) Our driverless dilemma. Science 352 (6293): 1514-1515.

Harris J (1975) The survival lottery. Philosophy 50: 81–87. doi:10.1017/s0031819100059118

Iltis A (2015) Aristotle's ethics and politics: reflections on bioethics and the contemporary state. History of Medicine 2 (4): 442–447.

Kortenkamp KV, Moore CF (2014) Ethics Under Uncertainty: The Morality and Appropriateness of Utilitarianism When Outcomes Are Uncertain. The American Journal of Psychology 127 (3): 367-382.

Kolber A (2009) The organ conscription trolley problem. Am J Bioeth 9 (8): 13-14. doi: 10.1080/15265160902948298.

Lally JF (2015) Ebola and moral philosophy: the trolley problem as a guide. Del Med J 87 (1): 25-26.

Manthous CA (2014) Emergency surgery, cardiopulmonary resuscitation, and the trolley problem. J Crit Care 29(1): 170-171. doi: 10.1016/j.jcrc.2013.10.007

Noothigattu R, Gaikwad SS, Awad E, Dsouza S, Rahwan I et al. (2017) A Voting-Based System for Ethical Decision Making. Published online: 20 September 2017. https://arxiv.org/abs/1709.06692?context=cs.AI.

Rai TS, Holyoak KJ (2010) Moral principles or consumer preferences? Alternative framings of the trolley problem. Cogn Sci 34 (2): 311-321. doi: 10.1111/j.1551-6709.2009.01088.x

Rosenberg A (1992) Contractarianism and the "trolley" problem. J Soc Philos 23 (3): 88-104. doi:10.1111/j.1467-9833.1992.tb00134.x

Samokhvalov A, Krylov N, Vychuzhanin D (2017) Sindrom emotsionalnogo vygoraniya u vrachey (skolko let mne ostalos?) [Burnout syndrome in physicians (How long shall I last?)]. Vrach [Doctor] 9: 2-5. (In Russ.)

Swann WB, Jr., Gómez A, Dovidio JF, Hart S, Jetten J (2010) Dying and killing for one's group: identity fusion moderates responses to intergroup versions of the trolley problem. Psychol Sci 21 (8): 1176-1183. doi:10.1177/0956797610376656

Thomson JJ (1976) Killing, letting die, and the trolley problem. The Monist 59 (2): 204-217.

Thomson JJ (1985) The Trolley Problem. Yale Law Journal 94 (6): 1395-1415.

Tversky A, Shafir E (1992) The Disjunction Effect in Choice Under Uncertainty. Psychological Science 3 (5): 305-309. doi:10.1111/j.1467-9280.1992.tb00678.x

About the authors

Nikolay Nikolaevich Krylov – Doctor of Medical Sciences, Professor, Department of Human Studies, Institute of Social Science, FSAEI HE I.M. Sechenov First MSMU MOH Russia (Sechenov University), Moscow. Email: nnkrylov01@yandex.ru

Yevgeniya Lvovna Panova – Candidate of Philosophical Sciences, Associate Professor, Department of Human Studies, Institute of Social Science, FSAEI HE I.M. Sechenov First MSMU MOH Russia (Sechenov University), Moscow. Email: evepanova@gmail.com

Aftandil Vagifovich Alekberzade – Doctor of Medical Sciences, Professor, Department of Surgery of the Institute of Clinical Medicine named after N.V. Sklifosovsky, FSAEI HE I.M. Sechenov First MSMU MOH Russia (Sechenov University), Moscow. Email: aftandil.v.alekberzade@gmail.com