GRAPH FORMALISM FOR NORMATIVE ETHICAL TASKS IN ARTIFICIAL INTELLIGENCE

BORIS D. GROZDANOFF

Institute of Philosophy and Sociology, Bulgarian Academy of Sciences

Defence and International Security Institute (DISI)

QAISEC

grozdanoff@gmail.com

Abstract

Within the diverse and rich debate around the so called "Ethical Artificial Intelligence" one of the major setbacks is the accepted lack of a unified ethical system that can be used for modelling the behavior of AI run systems like artificial moral advisors (Savulescu and Maslen 2015), autonomous vehicles, autonomous weapons and robots. Following upon an earlier suggestion for an artificial moral agent (2020) I will suggest a more general method for the quantification of key ethical concepts like consequence, behavior, rule, justification, state and property. The method would allow to harness the enormous computational power of modern computers for the purposes of in-depth ethical analysis of normative non-trivial cases. If successful it would provide a transmission mechanism, susceptible to rational evaluation, between ethics as a traditional discipline from the field of humanities and computation, as a mathematical structure of the kind found in natural sciences, where it provides not merely means of rigorous formulation but also of testing, evaluation, proofs and growth of scientific knowledge. The method introduces numerical weights assignment on the basis of ethical non-identity. Thus, if between a certain ethical dilemma with two possible actions we observe that the actions are normatively non-identical we can continue and capture this non-identity in graph-numerical terms thus effectively quantifying the normative setting. This quantification would provide us with the rigorously formalized content that would allow us to transfer this rigor onto a subsequent logical reasoning, enabling inferences, preference, justification and choice.

Keywords: ethical artificial intelligence, AI, ethics, artificial moral agent, ethical consequences, graphs, justification

Contemporary society is facing a *par excellence* two-culture [1] collision: the technological layer of global society is facing the inevitability of the unresolved problems of Western ethics along the echo of the ongoing explosion of artificial intelligence (AI). As the ubiquitous digital technologies evolve towards enormous volumes of data (designated today by the familiar Big Data term), the imminent problem of its employment, both in terms of usability

and control, finds its twin-answer in the non-trivial solution of artificial intelligence. The main effect of this collision is focused in the blooming research and effective deployment of artificially intelligent solutions, which act ethically and – in the greater number of cases – highly non-trivially so. From autonomous cars, street cameras, facial recognition systems and the Chinese social credit system [2] to defense systems such as autonomous weapons and ultra short notice early warning systems, AI-based systems do not merely raise ethical questions – they already *act* as ethical agents in the real life human-technology field. It therefore seems late and to some extent *ex post* to debate the ethical consequences of AI systems: that should have been done *before* their operational introduction. But, as we only all too well know, academic discussions, such as debates in ethics (as they figure within the Western philosophical tradition) hardly have the industrial chance to be taken into consideration when big corporations and states invest billions of dollars to achieve via artificial intelligence extremely specific tasks. And if some of their effects, direct or indirect, happen to have perceived as a positive ethical value – only the better. But the cold reality of operating artificially intelligent systems is such that there is no universally agreed upon systematic ethics which, first, to be taken into account at the stage of developing technologies; and second, within the enormous number of such candidates, every one receives its identity by rivaling most of the others.

The main challenges before ethics of artificial intelligence are the lack of a universally accepted moral system [3], the lack of universally accepted rationally justified moral criteria, and the technical but non-trivial challenge of *how to embed* any sort of ethics in real time operating artificially intelligent systems. Thus, contemporary society is facing a practical dilemma that needs a demonstrably successful solution in the conditions of a severe *zeitnot*. We should either continue to observe the exploding operational deployment of systems run by artificial intelligence and remain content with all immediate ethical effects. The ethics in this stance is not the ethics we discover, formulate, justify and embed into the systems. Instead, the ethics in this stance is the blind result of observed actual effects of operating artificial intelligence.

Or, we should try to develop a better solution to the "observed" ethics. A solution that would be demonstrably superior than the blind normativity artificially taking place. In the ideal scenario we need to discover, justify and develop an integrated ethical system of standards, carried by rules, principles, actions and ethical agents. This system should be demonstrably

better than its alternatives, and – most of all – better than the "observed" alternative. And its structure should also be *technologically compatible* with existing and future systems of artificial intelligence. That is, it needs to be quantified in formulation so that is embeddable in modern computational devices, from solitary chips to super computers, all of which, naturally, have a computational and thus mathematically-logical architecture.

The main method of superiority demonstrability is the critical comparison with alternate solutions. For example, if a quantified normative system (QNS)[4] recommends a certain action to the operating system of an autonomous vehicle, this recommendation should be demonstrably superior with respect to any other action within the system and also outside the system. An outside QNS ethical solution is such that is arrived at by an alternative system of ethics and both solutions are available to undergo rational comparison, evaluation and order by superiority.

Between the general branches of Western ethics, the deontological approach and the normative approach, I chose – for purely practical reasons – the latter. In normative ethics the focus is on arriving at, formulating and justifying moral criteria and rules for right and wrong. Effectively, normative ethics is attempting to rationally discover *how* the foundation of our ethics emerges. Its main tools are discovery and rational justification.

The general deontological approach in ethics is focused on moral rightness as a kind of a property inherently possessed (or not) by bearers like principles and actions. From a practical perspective, such as the one posed by our current task, the primary goal is to effectively arrive at morally superior principles, rules and actions in nothing short of demonstrable way. The instruments of quantification and logical reasoning would provide rational and thus demonstrable ways of arriving at morally superior rules, actions and principles.

In an influential paper from 2007, Anderson and Anderson argued for the possibility to create a machine that functions successfully as an "explicit ethical agent" [5] and their demonstration had the discreet structure of 6 steps. In 2008, famously, Wallach and Allen[6] provided what is arguably the most influential discussion of artificial moral agents, within the emerging field of roboet hics. The formulation of the *Moral Turing Test* by Allen, Varner, and Zinser in 2000 provided the first rigorous method of assessing the ethical success of an artificial moral agent (AMA). Powers (2006) suggested a logical (though not a computational, *sic!*) framework to arrive at a Kantian moral machine, arguing that a machine–computible (version

of the) Kantian categorical imperative can be based on (a certain choice of) logic. In 2010 Beavers pointed to the "reason machine ethics cannot move forward in the wake of unsettled questions such as", as he dubs it, "the hard problem of ethics", which asks what it is that effectively makes moral behavior (be it that of a human or that of a machine) moral in the first place.

The debate on the possibility and nature of artificial moral agents gains traction mainly in two directions. The first one is governed by *theoretical focus*, perhaps best captured in Beavers's "hard problem". The other one is the *practical direction*, illustrated by suggestions like that of Powers for a Kantian machine. Following the second general focus, but, unlike Powers, along the computational line rather than the purely logical line, I would like to suggest a technical proposal, which is exclusively practical and describes a formal scaffolding for an AMA without actually developing one – that would be a heavy, detailed and computational task to be developed further. The proposal suggests that we can attempt to use (an enhanced version of) graph formalism in order to formalize primitive terms from normative ethics and thus allow for quantification over them and any resulting relationships, on the one hand, and, on the other, to reason logically on the basis of the resulting graph and thus of computational structures.

At the core of the project is the instrument of graphs, as found in classical graph theory [7], which, I will argue, can be used as a language for formalizing ethical problems, on the one hand, and, on the other, can be used as a rational tool, *both computational and logical*, for preparing and justifying solutions for them.

The practical use of this proposal would become truly available for rational assessment, both ethical and computational, only after a functional prototype model is actually developed and tested in ethical scenarios that mock real life. In the beginning, the space of ethical possibilities would be minimal, such as "run over a pedestrian" or "do not run over a pedestrian" in an autonomous vehicles setting; and gradually the GNF space would be enriched so that it may become expressive and powerful enough to quantify a multitude of real life scenarios. Once developed and trained, the GNF-based models would provide us with a rationally assessible basis for moral criteria. They would also be fully comparable to any applicable alternative models. Where they would di ffer positively is their rigor, complexity, transparency, explainability and computational and logical canvas.

The anticipated benefits of this model (dubbed here Graph Normative Formalism or GNF) would be, on the theoretical side, its systematized version of real-life ethical choices and justification along with the heuristics of the ethical model that emerged in formalism; and on the practical side – its complete compatibility and integrability with existing legacy computational systems which operate even today and perform a myriad of ethically significant decisions as I write this. Autonomous vehicles, social credit systems, autonomous weapons and many others deployed as operational in the world, effectively already operate as artificial moral agents, whether we like it or not and whether we realize our being late or not. Human technology needed a far superior alternative to its artificially intelligent systems yesterday, not merely today, and the possibility to systematically embed a superior ethical solution into legacy machines is anything but trivial.

GRAPH NORMATIVE FORMALISM (GNF)

A *normative graph* (G_N) is a particular kind of data structure which is capable of modelling normative ethical relationships via the familiar node-edge structure from graph theory. For our purposes (G_N) takes the familiar definition of a graph as (in our case) a non-null set of nodes connected by edges. In its simplest form, the graph is a pair (N, E) where N is a set of nodes and E is a set of edges between the nodes, which express a binary relation between any two nodes in the pair [8]. Since in ethics scenarios the agency is a non-trivial action of a primitive kind, the graphs should be endowed with the ability to express the difference between an (ethical) agent and the subject of her action. This difference is best captured by the graph property of *directionality:* an action performed by agent A towards agent B would be expressed by a *direction* of the edge from A to B. The space of (G_N) manages to express normative information, to order it, to reflect the order observed in real life scenarios, and to connect all agents and their actions with all normatively significant effects within this space. But most of all, it would allow us to *quantify normative agents, actions and effects,* and to reason about them on the basis of this quantification.

The richness of the ethical scenarios we find in real life requires that our graph formalism be able to express edges' branching as nodes and, in addition to our standard binary relation, introduce the functionality of (undirected) *hyperedges*. A graph whose nodes are connected by a hyperedge is introduced here as a *hypergraph*. Due to the same reason we need

to introduce in the normative graph space the possibility to connect the same pair of nodes by a multitude of edges; consequently we will introduce the *multigraph* in the space. We will also introduce *loops* and *self-loops* in the space.

Thus, the graph normative formalism would provide the mathematical and logical structure necessary for applying the two most rigorous instruments of human rationality – *mathematics* and *logic* – to the field of ethics, which has so far been grasped solely as a humanities problem.

The formalism needs to adequately respond to a rich variety of real life ethical tasks. A non-exhaustive list should necessarily include primitive normative terms like:

- Agents – performers of ethically significant actions (A_N)
- Actions – the activities of ethical agents which have ethical significance (O_N)
- Properties (P_N)
- Events - events which have an ethical significance (E_N)
- Environment – the environment of normative actions and events where ethical agents operate and reason. The environment can be ethically loaded or ethically neutral (V_N)
- States – an ethically significant ordered relationship built up by agents, actions and events (S_N)
- History – States (H_N)

Generally, agents are represented by nodes, individual or connected in an ordered set within a graph. Actions, being of operational nature, are represented by edges. Edges also represent ethical properties such as morally right or wrong. Moral properties can be carried by agents, their actions and ethically significant events. Thus they are represented in GNF by nodes, edges (hyperedges) and graphs (hypergraphs). They can also be carried by groups of nodes, edges (hyperedges) and graphs (hypergraphs).

Normative relationships are expressed by the complete functionality of GNF, which includes directionality of edges, number of edges connecting a node, normative weights of an edge (assessing the normative weight of actions, expressed by the edge), of a node (as the sum of all weights of all edges directed outbound of the node) and of graphs (as both kinds of sums).

Another type of normatively significant phenomena that can be expressed in GNF are the phenomena of *normative traceability*. Thus, a normatively significant event taking place when agent A acts affecting agent B and expressed by a graph (G) of two nodes connected by

a directed edge from A to B can be traced to all other actions and graphs that are at least partially resulting of it. GNF provides a sufficiently rich and expressive space to formalize and dynamically change a variety of normative histories. Normative traceability can provide the complete formal picture of the sets of all effects of any ethical agent's actions.

Passive actions, which are so important in normative ethics, can be expressed in GNF through a bypassing graph via a complementary bypassing node that leads to the negation of the event of the agent's being passive.

The GNF can serve in a multitude of non-trivial ways for our normatively ethical analysis:

- It can formalize normatively significant situations
- It can serve to integrate them into a common analyzable space
- It can assess the structured ethical weights of agents, events and actions.
- It can compare agents, events and actions on the basis of available information
- It can prove theorems in a formalized normative space
- It can formulate conjectures in the normative space that can be researched and eventually decided upon rigorously (by proving them in a theorem or denying them, also in a theorem)
- It can provide significant reasons to recommend certain actions and structured sets of actions over others on the basis of available information in the relevant formal GNF space
- It can provide normative justification of an agent's action and compare it to actual or potential alternative actions
- It can extract explainability and transparency from histories of actions and, especially usefully so, from very complex sets of them.

The awareness of an action and its consequences is arguably key for the ethical assessment of an action so our GNF would be ethically incomplete without epistemic functionality. The simplest way to integrate it into GNF is to add *awareness* and *intention* as weighed parallel edges to the main edges, allowed by the multigraph structure we have already defined on GNF. Thus, in the case where A kills B via the action *a,* there would be a_a and a_i parallel edges to *a* with their own weights. The quantified integration of the indices' values can be approached either by multiplication or by summation; however, multiplication seems much more apt for the purposes of arriving at a lesser number of values assigned to a certain action.

A problem might arise when we try to multiply a weight by zero, for example, the zero here being the value of the awareness index. In standard arithmetic multiplication the end result should be zero, but in our case that would amount to annihilation of the alleged non-zero value of the main index of the action *a*. Therefore, it seems best to express the indices' values by their order in a set of a tensor type, which would allow us an immense functionality between the separate indices of simple graphs and their relationships with larger graphs, and also, the transformation of a graph into its historical successor. Thus, the tensor calculus here might turn out to be a proper candidate for expressing any number of necessary indices, allowing for tensors of different shapes and dimensionality. The tasks of tensor operations between tensors of different shapes and dimensionality can be approached, for example, by using Python's programming language NumPy module [9].

Three points need to be stressed in particular. First, since every agent, action and event of a normative significance are connected in real life with a great number of others, the richer the GNF space – the richer the formal facts, and more importantly – the more rigorous the facts. Let us present an example with the infamous *Trolley Problem* [10] – it is one of the perhaps best known illustrations of extreme non-triviality of a normatively significant environment. As we typically encounter the TP in literature, we begin with a simple and even formal version of it, where its form has a significant priority over its content, and only then do we continue to add content in the form of different normative scenarios, where the solution of the form of the TP leans more towards one direction or the other, depending on the different content assigned.

In reality, however, things are never formal – they are always *contenfully actual*. What can be taken as formal in real life scenarios is not the objective lack of content that has the rational potential to solve the dilemma (or similar dilemmas) with a multitude of respective rational justifications for or against each choice. It is rather *the lack of sufficient epistemic access to the actual ethical content* that is the best candidate for an objective normative decision and action recommendation. It is therefore of crucial importance to introduce an indexical weight of richness for agents, actions, events and operational normative space in general to our GNF. Naturally, the smaller the index value, the less rationally reliable any theorem on GNF should be considered. Thus, this index can be assigned to any conjecture or theorem in GNF, and that way it can provide us with additional normatively relevant information that should be a factor in normative decisions.

The second point is the *isomorphism between computational graph space and logical graph space* where we can use one space to serve both computational tasks (quantifying over weights, their sums and over graph structures) and logical tasks, where we can endow the same GNF with a logical layer expressing, provisionally, a 1st order predicate non-monotonic multimodal logic, rich enough to deliver a significant number of inferences of a normative nature.

The third point is the integrability and trainability of GNF. GNF-based systems can be easily integrated into legacy computational devices. But also, and much more importantly, they can be fed with existing databases and trained via neural network models which can build up an AI system capable of demonstrably superior normative decisions.

At the end, the GNF models would allow for a rational order within a variety of ethical scenarios or choices on the basis not merely of quantified and related logical order, but also of richness or depth of GNF space. Thus, once again using the Trolley Problem as an illustration, if we take two extreme scenarios – in the first of which we have nameless people characterized by complete ethical inertness in terms of ethically significant content available in order to rationally justify a decision for a horn in the dilemma, and in the second of which we have a full blooded actual case with greatly detailed, ethically significant data available (for example, to give a somewhat easier setting – the single tied person is an innocent child, while the 5 other persons are runaway Nazi war criminals), we can use GNF to order them rationally. A significant advantage of the order would be the depth of ethically significant information available and formalized within the GNF space. This available depth can be formalized, in its own turn, via an additional depth-index equal to, for example, the connectedness rank of the node or graph participating, and in these two particular cases of the Trolley Problem the depth-index difference would be highly non-trivial.

A GNF model can formalize and integrate ethically significant data into a mathematical system, and can also reveal actual relationships within it. It can provide a rational basis for ethical analysis and create a future forecast for ethically significant behavior of agents such as AMAs. It can also provide a historical dynamic, embedded into its graph structure, for example with the introduction of a temporal precedence order via the hypergraph functionality allowed on it. A last but critical benefit for GNF models is their *AI trainability* on existing data. In one such scenario, described elsewhere [11], an AI system can generate ethically significant data,

say, extracted from real life scenarios, and train a GNF model so that it may learn the optimal ethical behavior within unprecedented massives of complexities, previously rationally inaccessible to humans.

NOTES

[1] As described by Snow in «The Two Cultures». For an in-depth discussion see the excellent book by Brown, J.R. «Who Rules in Science».

[2] For an intro see "How China Is Using Big Data to Create a Social Credit Score". – In: *Time*. (2019). Retrieved 21, November 2021.

[3] See Serafimova (2020) and Todorova's new book on Artificial Intelligence and its ethical aspects (2020).

[4] I suggest a QNS model in Grozdanoff (2020).

[5] Anderson, M., and S. Anderson. (2007). Machine ethics: Creating an ethical intelligent agent. – In: *AI Magazine* 28 (4): 15 – 26 . p. 22.

[6] See the enormously influential roboethics book *Moral Machines* (2009), especially chapters 2 and 4.

[7] For an introduction to modern Graph Theory see Benjamin, Chartrand and Ping's *The Fascinating World of Graph Theory*, 2017, Princeton University Press.

[8] In a consequantialist setting where not just agents and their actions are of interest, but also the effects of their actions, which will go on to quantify over via our (G_N), it is better to allow the binary relation between nodes to range over any two nodes in the space, for this would not restrict the connectivity and traceability of any agent and action to any effect. This ability of our formalism is highly desirable, due to the expressivity of real life ethics scenarios.

[9] For an intro to Python's NumPy see *Scientific Computing with Python: High-performance scientific computing with NumPy, SciPy, and pandas*, 2nd Edition, by Führer, Solem and Verdier (2021), Packt Publishing.

[10] Jarvis Thomson, Judith (1985). "The Trolley Problem". Yale Law Journal. 94 (6): 1395–1415.

[11] See Grozdanoff (2020).

REFERENCES

Allen, C., G. Varner, and J. Zinser. (2000). Prolegomena to any future artificial moral agent. – In: *Journal of Experimental & Theoretical Artifi cial Intelligence*. 12 (3), 251 – 261.

Anderson, S. (2011). How machines might help us to achieve breakthroughs in ethical theory and inspire us to behave better. – In: *Machine Ethics, ed. Michael Anderson and Susan Anderson*. New York, Cambridge University Press, 524 – 530.

Anderson, M., S. Anderson. (2007). Machine ethics: Creating an ethical intelligent agent. – In: *AI Magazine* 28 (4), 15 – 26.

Beavers, A. (2001). Kant and the problem of ethical metaphysics. – In: *Philosophy in the Contemporary World*, 7 (2), 47 – 56.

Beavers, A. (Ed). (2010). Robot ethics and human ethics. – In: *Special issue of Ethics and Information Technology*, 12 (3).

Benjamin, Chartrand and Ping. (2017). *The Fascinating World of Graph Theory*. Princeton, University Press.

Brown, J. R. (2005). *Who Rules in Science?* Harvard University Press.

Dennett, D. (1984). Cognitive wheels: The frame problem in artifi cial intelligence. – In: *Minds, machines, and evolution: Philosophical studies*, ed. C. Hookway. New York, Cambridge University Press, 129 – 151.

Dennett, D. (2006). Computers as prostheses for the imagination. Invited talk presented at the International Computers and Philosophy Conference, May 3, Laval, France.

Floridi, L., and J. Sanders. (2004). On the morality of artificial agents. Minds and Machines 14(3), 349 – 379.

Gips, J. (1995). Towards the ethical robot. – In: *Android Epistemtology*, ed. K. Ford , C. Glymour, and P. Hayes. Cambridge, MA: MIT Press, 243 – 252.

Grozdanoff, B. D. (2020). Prospects for a Computational Approach to Savulescu's Artificial Moral Advisor. – In: *Ethical Studies*, vol. 5, book 3, 121–140.

Kant, I. [1785] (1981). *Grounding for the Metaphysics of Morals,* trans. J. W. Ellington. Indianapolis, IN: Hackett Publishing Company.

Kurzweil, R. (1990). *The Age of Intelligent Machines.* Cambridge, MA: MIT Press.

Mill, J. S. [1861] (1979). *Utilitarianism.* Indianapolis, IN: Hackett Publishing Company.

Mittelstadt, B. D. et al. (2018). *The ethics of algorithms: Mapping the debate*. – In: *Big Data and Society*. Vol. 3 issue 2.

Moor, J. (2006). The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems 1541 – 1672, 18 – 21.

Plato. (1993). *Republic,* trans. R. Waterfi eld. Oxford, UK: Oxford University Press.

Powers, T. (2006). Prospects for a Kantian machine. – In: *IEEE Intelligent Systems,* 1541 – 1672, 46 – 51.

Ruse, M. (1995). *Evolutionary Naturalism.* New York, Routledge.

Savulescu, J., H. Maslen. (2015). Moral Enhancement and Artificial Intelligence: Moral AI*?* – In: Romportl J., E. Zackova, J. Kelemen (Eds). *Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics,* vol. 9, Springer, Cham.

Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. – In: *Humanities and Social Sciences Communications,* 7, 119.

Snow, Ch. P. (2001). [1959]. *The Two Cultures*. London, Cambridge University Press, 3.

Todorova, M. (2020). *The Artificial Intelligence: A Brief history of its development and the ethical aspects of the problem,* Sofia, East-West Publishing, in Bulgarian.

Turing, A. (1950). Computing machinery and intelligence. *Mind* 59 (236), 433 – 460.

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. – In: *Ethics and Information Technology,*12 (3): 243 – 250.

Wallach, W., and C. Allen. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York, Oxford University Press.