Neutrosophic Sets and Systems, Vol. 75, 2025
University of New Mexico

Introduction to algorithmic ethics under a neutrosophic framework: Re-mapping the debate

Antonios Paraskevas [1,*] and Florentin Smarandache [2]

[1]University of Macedonia, Department of Applied Informatics, 156, Egnatia Str., 54636, Thessaloniki, Greece; aparaskevas@uom.edu.gr;
[2]University of New Mexico, Mathematics Department, Gallup, NM 87301, USA; smarand@unm.edu;

* Correspondence: aparaskevas@uom.edu.gr;

Abstract: Algorithms have become an integral part of professional life and social interactions, in health, transport, commerce and industry. Moreover, they are bringing about changes in the natural, social and human sciences, enriching our knowledge and testing the limits of technology. In the age of artificial intelligence (AI), the ethical behavior of algorithms is facing increasing scrutiny. Traditional models that attempt to evaluate algorithm ethics using binary or probabilistic methods often fall short in addressing the complexity and uncertainties present in real-world situations. This study introduces a novel approach to assessing algorithmic ethics by utilizing neutrosophic indeterminacy. Neutrosophic logic, which considers three parameters - truth (T), falsity (F), and indeterminacy (I) - provides a robust framework for capturing the ambiguity and inconsistencies that arise in ethical decision-making processes. By utilizing this approach, we attempt to develop a method for measuring ethical ambiguity in algorithmic decision-making. By applying the proposed conceptual framework to an illustrative example we highlight the capacity of neutrosophic logic to capture and measure ethical uncertainties in a more comprehensive manner, thus offering a new tool for evaluating the ethical integrity of algorithms in complex environments.

Keywords: Neutrosophic ethical integrity score; neutrosophic logic; algorithmic ethics; fairness; transparency; accountability.

1. Introduction

Algorithms subtly influence our lives in various ways. Operations, judgements, and choices that were previously left to people are increasingly being delegated to algorithms, which may advise, if not determine, how data should be evaluated and what actions should be performed in response. Examples are abundant. Profiling and categorization algorithms influence how people and groups are formed and managed [1]. Recommendation systems guide users on when and how to exercise, what to buy, which route to take, and who to contact [2]. Data mining techniques are believed to have promise for making sense of growing streams of behavioral data provided by the 'Internet of Things' [3]. Personalization and filtering algorithms continue to be used by online service providers to manage information access [4-5]. Machine learning algorithms automatically detect misleading, biased, or erroneous knowledge at the time of generation.

In this paper we adapt the following definition about an algorithm: "a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions" [6].

In the broader context, the rapid expansion of algorithms in crucial industries such as healthcare, banking, and law enforcement has raised significant ethical concerns. Algorithms have the potential to worsen biases, diminish transparency, and weaken accountability. There are many factors that make it difficult to determine the potential and actual ethical influence of an algorithm. Identifying the role of human subjectivity in algorithm design and configuration often requires an examination of long-term, multi-user development processes. Learning algorithms, which are frequently cited as the "future" of algorithms and analytics [7], add uncertainty into how and why judgements are made because of their ability to change operational parameters and decision-making rules "in the wild" [8]. The potential for algorithms to improve individual and social welfare comes with significant ethical risks [9]. In conclusion, algorithms are not ethically neutral.

In recent years, researchers have attempted to identify and categorize the ethical problems that algorithms give rise to and the solutions that have been proposed in recent relevant literature. The conceptual map proposed by [10] (Fig.1) remains a fruitful framework for reviewing the current debate on the ethics of algorithms.

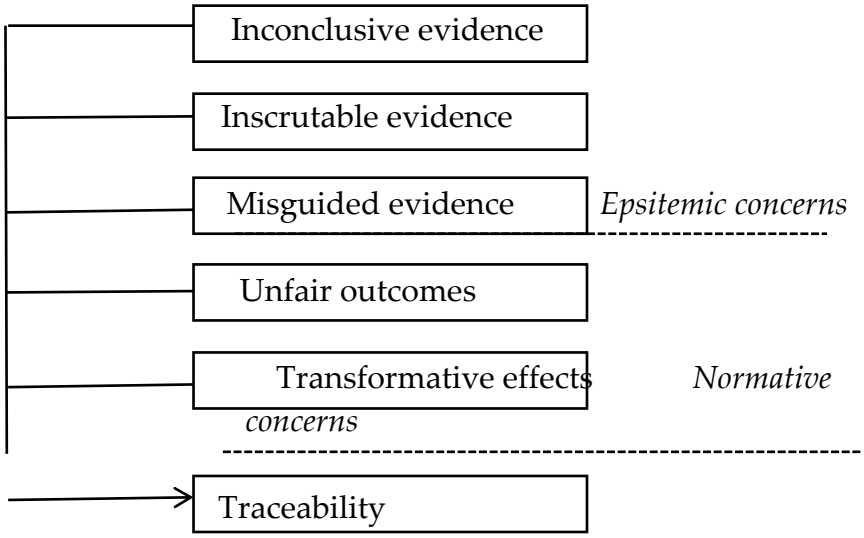

Figure 1. Six types of ethical concerns raised by algorithms [10]

Inconclusive evidence refers to the inevitably uncertain knowledge when algorithms draw conclusions from the data they process using inferential statistics and/or machine learning techniques.

Inscrutable evidence refers to the accessibility between the data and the conclusion in the case data are used as (or processed to produce) evidence for a conclusion.

Misguided evidence means that conclusions can only be as reliable (but also as neutral) as the data they are based on.

Unfair outcomes refers to the observer-dependent 'fairness' of the action and its effects.

Transformative effects denotes the unintended and significant changes that algorithms can impose on individuals or society, such as altering behaviors, decision-making processes, or social structures, often without transparency or user consent.

Traceability implies that harm caused by algorithmic activity is hard to debug (i.e. to detect the harm and find its cause), but also that it is rarely straightforward to identify who should be held responsible for the harm caused.

Driven by the definition of inconclusive evidence and its observed uncertainty, we are re-mapping the debate by introducing the concept of indeterminacy into all aspects of the ethical map (Fig.1). This approach strengthens our ability to address the complexity and ambiguity inherent in ethical decision-making for algorithms. By providing a structured approach to assessing how uncertainty affects ethical decisions, we increase accountability and transparency. This research

introduces a new neutrosophic methodological framework for evaluating ethical correctness (truth), ethical breaches (falsehood), and uncertainty (indeterminacy) in algorithms, offering a more comprehensive approach to ethical assessment. Thus, the conceptual idea of our study is depicted in Figure 2 and is posed as an organizing structure that allows a neutrosophic rigorous diagnosis of ethical challenges related to the use of algorithms.

Furthermore, aligning the ethical evaluation process with the intricacies of real-world scenarios makes the framework more useful and effective for guiding responsible and informed algorithm design and implementation. Introducing an "ethical score" metric that deals with indeterminacy can help tackle problematic situations such as ambiguous outcomes, data quality and representation, complex interactions, conflicting ethical principles, and human interpretation.

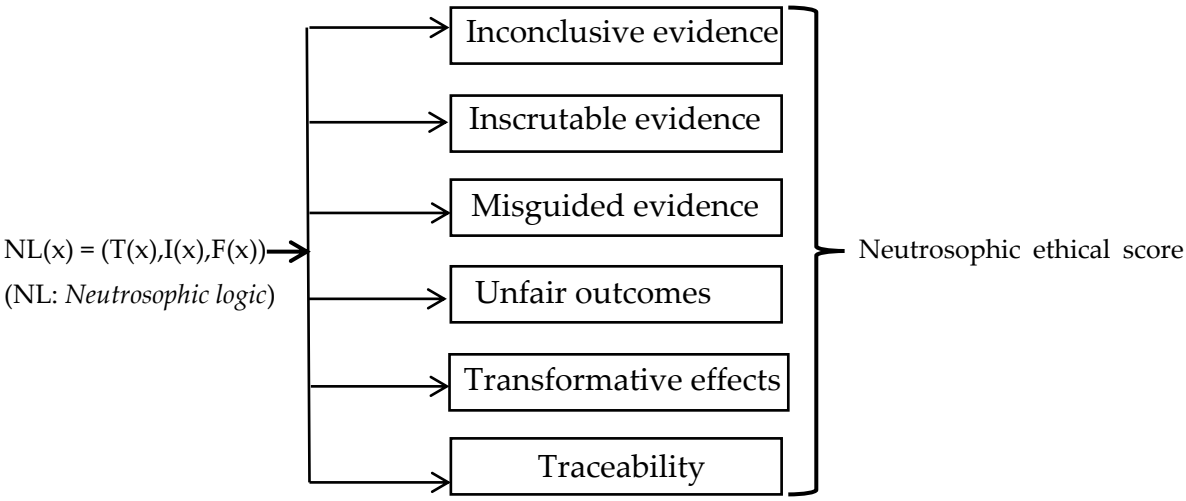

Figure 2. Neutrosophic algorithmic ethical map

Our review of the literature has revealed no prior studies that utilize this approach to systematically evaluate ethical ambiguity in algorithms, particularly in scenarios where principles like fairness, transparency, and accountability are in conflict. For this reason, this paper examines the significance of indeterminacy in ethical evaluations of algorithms and explores how a neutrosophic approach could impact the design of future algorithms. Actually, in our study, we aim to map out the ethical problems that arise from algorithmic decision-making by posing two main questions: 1) how can we systematically evaluate ethical uncertainty in algorithmic decision-making using neutrosophic logic? and 2) how does indeterminacy manifest in the conflicting ethical principles of justice, transparency, and accountability in algorithmic design?

While significant emphasis has been placed on concerns regarding prejudice, fairness, and openness [10-13], existing ethical frameworks sometimes struggle to navigate situations of uncertainty or conflicting ethical principles. In this way, our study also pinpoints the overlooked concept of indeterminacy and offers strategies for identifying ethical dilemmas, addressing a gap in the literature where traditional binary and fuzzy logic methods may fall short.

The structure of this article is as follows: In Section 2, we define and explain the neutrosophic mathematical framework needed to "construct" our proposed algorithmic “ethical index”, namely the Overall Neutrosophic Ethical Integrity Score (OvNEIS). This score will help us quantify the ethical performance of an algorithm in a neutrosophic environment. Next, in Section 3, we highlight the applicability of the suggested OvNEIS in an illustrative example from the field of healthcare. We examine an algorithmic decision-making process under three ethical criteria: fairness, transparency, and accountability and we briefly comment on the obtained results. Lastly, the "Concluding Remarks" section wraps up the key points of our study and proposes potential research work.

2. Materials and Methods

In neutrosophic logic, a concept A is T% true, I% indeterminate, and F% false, with (T, I, F) ⊂]]-0, 1+[[^{3}, where]]-0, 1+[[is an interval of hyperreals.

In this model, truth, falsehood, and indeterminacy may coexist, allowing for a more comprehensive representation of complex and ambiguous information. Sets containing neutrosophic components are employed in neutrosophic logic, with constituents having degrees of truth, falsehood, and indeterminacy. Its capacity to deal with ambiguity and uncertainty makes it useful in circumstances where standard logic systems may fail to offer correct representations.

In this framework, a formula φ is characterized by a triplet of truth-values, called the neutrosophical value defined as [14]:

NL(φ) = (T(φ), I(φ), F(φ)) where (T(φ), I(φ), F(φ)) ⊂]]-0, 1+[[^{3} (1)

For each algorithmic ethical criterion C_i $(i = 1,2 \dots, n)$, we can represent its performance by utilizing the concept of neutrosophic set in a similar way as given in (1):

$$C_i = \{(x, T_i(x), I_i(x), F_i(x)) | x \in U\} \quad (2)$$

where U : universe of discourse (e.g. algorithm outputs)

Next we define a formula, namely Neutrosophic Ethical Integrity Score (NEIS), which will be utilized for evaluating the ethical performance of algorithms based on the principles of neutrosophic logic. It is designed in such a way so as to balance positive and negative aspects (subtraction of falsehood) and capture the complexity of ethical evaluation (indeterminacy as a modifier)

Definition 1. The Neutrosophic Ethical Integrity Score (NEIS) for a given ethical criterion $\boldsymbol{i}$ is defined as:

$NEIS_i : \mathbb{R}^3 \rightarrow \mathbb{R}$

where $\mathbb{R}^3$ represents the three-dimensional space of the components of ethical evaluation, specifically truth, falsehood, and indeterminacy.

The $\boldsymbol{NEIS_i}$ function is given as follows:

$$NEIS_i = T_i\text{-}F_i\text{+}I_i \quad (3)$$

where :

$T_i \in [0,1]$ represents the degree to which the ethical criterion is satisfied by the algorithm. It is obvious that a value of T_i = 1 means complete satisfaction and T_i = 0 indicates complete dissatisfaction.

Following the same logic, F_i = 1 indicates complete violation to ethical criterion i and F_i = 0 means no violation.

Extra care should be given when evaluating the value of I_i. In our context, I_i= 1 indicates complete indeterminacy (uncertainty or ambiguity) in the ethical evaluation, while I_i= 0 means no indeterminacy.

Once the individual NEIS values are calculated for each criterion, the overall NEIS can be computed. This may involve taking a weighted average or simple average of the individual scores, depending on the specific context and importance of each criterion.

Definition 2. Let n be the number of criteria for assessing algorithmic ethics. The Overall NEIS can be defined as:

$$\text{Overall NEIS (OvNEIS)} = \sum_{i=1}^{n} w_i * NEIS_i \quad (4)$$

where w_i is the weight assigned to algorithmic ethical criterion i such as $\sum_{i=1}^{n} w_i = 1$.

3. Results

Let us now study an illustrative example that applies the proposed Neutrosophic Ethical Integrity Score (NEIS) to an algorithmic decision-making process in the context of healthcare. Let us assume that we wish to measure the performance of an algorithm designed to recommend treatment plans for patients in the following three ethical criteria: (i) *fairness*, (ii) *transparency*, and (iii) *accountability*.

To assess its ethical integrity, we evaluate its performance using the three aforesaid criteria in the next framework:

(i) Fairness: Does the algorithm treat all patients similarly, regardless of demographics (such as age, gender, or ethnicity)?

(ii) Transparency: Are healthcare practitioners able to comprehend the algorithm's decision-making process?

(iii)Accountability: Can the algorithm's outputs and decisions be tracked back to the responsible stakeholders?

Next, we will calculate the NEIS for each criterion.

Let us assume that based on historical data or surveys and expert evaluations we have the following assignment of values for each criterion (Equation 2):

(i) Fairness: C_f = (0.8, 0.1, 0.1)

(ii) Transparency: C_t = (0.7, 0.1, 0.2)

(iii) Accountability: C_a = (0.6, 0.1, 0.3)

The NEIS for each criterion is calculated using Equation (3):

$NEIS_{C_f} = T_{C_f} - F_{C_f} + I_{C_f} = 0.8$

In the same way, we get :

$NEIS_{C_t} = 0.6$ and

$NEIS_{C_a} = 0.4$

Let us know make the fair assumption that the weights of each ethical criterion is equal, i.e. $w_{C_f} = w_{C_t} = w_{C_a} = 0.33$.

Remark: depending on the context of the problem we are studying, different weights could be assigned to the criteria. This could be useful in cases where the fairness criteria is considered more important than the accountability and transparency criteria (e.g. hiring algorithms, loan approval algorithms, etc.). Therefore, a higher weight value for the fairness criterion will be assigned.

In our scenario, the overall NEIS is calculated by applying Equation (4):

OvNEIS = $w_{C_f} * NEIS_{C_f} + w_{C_t} * NEIS_{C_t} + w_{C_a} * NEIS_{C_a} = 0.6$

The Overall NEIS of 0.6 reflects an average ethical performance of the algorithm across all three criteria (Fairness, Transparency, and Accountability), given that all criteria are equally important.

A NEIS of 0.6 indicates moderate ethical integrity, with some space for improvement, notably in criteria such as accountability, which scored lower (0.4) compared to ther ethical criteria.

Furthermore, we can make the following observations based on our proposed method:

- The high fairness score indicates that the algorithm is unlikely to perpetuate considerable prejudice or discrimination, a critical worry in many algorithmic systems.
- The algorithm's decision-making process is only partly transparent, weakening confidence and making it difficult for consumers to comprehend or question its suggestions.
- The low accountability score raises questions about who is responsible for the algorithm's results, particularly in times of error or harm.

Next steps for the improvement of the algorithm:

- Continue to check for prejudice and fairness concerns, particularly in circumstances where indeterminacy is high.
- Improve the algorithm's capacity to explain its decision-making process to users, particularly when the decision is not clear.
- Involve users (e.g., healthcare experts) in assessing the algorithm's explanations to ensure that transparency has significance.
- Implement organizational policies that define accountability for algorithmic judgements, especially when they may cause harm.

4. Conclusions

Since 2016, the ethics of algorithms has been a major issue of debate among academics, technology providers, and regulators. The topic also gained pace as a result of the so-called "summer of AI" and the widespread deployment of Machine Learning algorithms. One aspect that was not explicitly captured by the research debate, and which is becoming a central point of discussion in the relevant literature, is the increasing focus on the use of algorithms, AI and digital technologies more broadly, to deliver socially good outcomes [15-17].

Starting from an admittedly simplistic notion of ethics as ''the study of what we ought to do,'' our aim has been to sketch a model that could "quantify" algorithmic ethics. In this perpserctive, we propose a formal mathematical neutrosophic framework that assists stakeholders in making key decisions based on a measurable indicator, OvNEIS. The latter, which is intended to assess uncertainty in ethical judgements in algorithmic decision-making, incorporates neutrosophic logic concepts such as truth, falsehood, and indeterminacy, allowing for a more comprehensive representation of complex and ambiguous data.

Specifically, OvEIS can address and add value, by considering the key role of indeterminacy, to the following issues concerning the algorithmic ethics:

- Inconclusive and Misguided Evidence: it helps to systematically quantify and address cases where evidence is either inconclusive or misleading.
- Inscrutable Evidence: it can enhance traceability and make algorithmic processes less inaccessible, answering concerns about traceability.
- Unfair Outcomes: it can analyse fairness as one of its primary aspects, guaranteeing that algorithms are more thoroughly reviewed for bias and fairness by taking into account not just truth and falsity, but also indeterminacy in their outcomes.
- Transformative Effects: it can consider the uncertainty in their long-term impacts on society, highlighting potential ethical concerns before they manifest.
- Traceability and Accountability: by scoring algorithms on this criterion, we ensure that ethical violations are not only discovered but also quantified, supporting responsible design and implementation.

Future study should focus on extending this technique to a broader range of case studies from a variety of sectors. Validation should be enhanced and a comparative study with other fuzzy/neutrosophic MCDM methods could shed light on issues such as efficiency, complexity and robustness of the proposed method.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, 25(4), 435–437.
2. de Vries, K. (2010). Identity, profiling algorithms and a world of ambient intelligence. *Ethics and Information Technology*, 12(1), 71–85.
3. Portmess, L., & Tower, S. (2014). Data barns, ambient intelligence and cloud computing: The tacit epistemology and linguistic representation of Big Data. *Ethics and Information Technology*, 17(1), 1–9.
4. Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems*, 24(1), 3–14.
5. Taddeo, M., & Floridi, L. (2015). The debate on the moral responsibilities of online service providers. *Science and Engineering Ethics*, 1–29.
6. Hill, R. K. (2016). What an algorithm is. *Philosophy & Technology*, 29, 35-59.
7. Tutt, A. (2016, March 1). An FDA for Algorithms. *69 Admin. L. Rev. 83 (2017)*, Available at SSRN: https://ssrn.com/abstract=2747994 or http://dx.doi.org/10.2139/ssrn.2747994.
8. Burrell, J. (2016). How the machine 'thinks:' Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1): 1–12.

9. Floridi, L., & Taddeo, M. (2016). What is data ethics?. Philosophical Transactions of the Royal Society A: Mathematical, *Physical and Engineering Sciences*, 374(2083), 20160360.
10. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2053951716679679.
11. Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency (pp. 149-159)*. PMLR.
12. Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values*, 41(1), 77-92.
13. Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive June*.
14. Smarandache, F. (1999). *A unifying field in Logics: Neutrosophic Logic*. In Philosophy, American Research Press, pp. 1–141.
15. Hager, G. D., Drobnis, A., Fang, F., Ghani, R., Greenwald, A., Lyons, T., & Tambe, M. (2019). Artificial intelligence for social good. *arXiv preprint arXiv:1901.05406*.
16. Paraman, P., & Anamalah, S. (2023). Ethical artificial intelligence framework for a good AI society: principles, opportunities and perils. *AI & SOCIETY*, 38(2), 595-611.
17. Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, 3(2), 111-115.

Received: June 27, 2024. Accepted: Oct 13, 2024