Towards Ethical Risk Assessment of Symbiotic AI Systems with Fuzzy Rules

Abeer Dyoub[1,*], Francesca Alessandra Lisi[1]

[1]University of Bari Aldo Moro, DiB Dept., via E. Orabona 4, Bari, 70125, Italy

Abstract

Artificial Intelligence (AI) based systems are expanding rapidly in all domains of life. They are entering our everyday life and performing tasks on our behalf. AI-based systems such as personal healthcare assistants are increasingly engaging in close symbiotic relationships with humans. Symbiotic AI (SAI) promises improved outcomes in various domains such as healthcare, education, and business. However, as the degree of symbiosis increases, so does the ethical risk. To ensure that these systems behave ethically and do not cause harm of any kind (physical, mental, violation of privacy, etc.), we need to find ways to assess the ethical risk (risk of causing harm), then choose the right action to mitigate that risk. In this work, we propose an approach based on fuzzy logic for ethical risk assessment (ERA) of SAI systems. The approach is illustrated by means of a case study taken from the healthcare domain.

Keywords

Symbiotic AI, AI Ethics, Ethical Risk Assessment, Fuzzy Logic, Fuzzy Rules

1. Introduction

Artificial Intelligence (AI) systems are rapidly expanding across all areas of life, becoming integral to our everyday activities and performing tasks on our behalf. These AI-based systems are increasingly forming close symbiotic relationships with humans, exemplified by digital twins, personal healthcare assistants, and virtual avatars. The term *symbiosis* has emerged as a result of the ongoing debate over whether AI will replace or enhance human abilities. Human-AI Symbiosis (known as Symbiotic AI, here after SAI) is about human-AI teaming, enabling people and AI to collaborate together for achieving better results together than they could separated [1]. SAI holds the promise of improved outcomes in various sectors such as healthcare, education, and business. However, SAI poses not only several technological challenges but also many philosophical questions. A deeply ingrained symbiosis guided by ethics will enhance human experience while respecting our values, which will make AI technologies ethically acceptable [2]. It is important to conceptualize and design holistic symbiotic frameworks for AI aiming at generating fair, legitimate, and effective outcomes while ensuring ethical and legal compliance. Such frameworks are expected to shape the development of SAI systems and influence technological governance through rigorous model assessment [3] .

Over decades, AI developers have been considering various moral or ethical theories for developing *artificial moral agents* (AMA). When people think about moral theories, they usually consider three primary schools of academic ethics that aim to explain what is good or bad, right or wrong, and why. Briefly, these three schools are [4]:

- Consequentialism: This theory, with utilitarianism as a notable example, asserts that the right action is the one that brings about the best overall consequences.
- Deontology: Exemplified by Immanuel Kant's theory, deontology states that an act is right or wrong based on its adherence to a set of principles, independent of its consequences. Breaking the rules is wrong, even if the outcome is positive.

3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE24), co-located with AIxIA 2024, November 25-28, 2024, Bolzano, Italy

[*] Corresponding author.

abeer.dyoub@uniba.it (A. Dyoub); francesca.lisi@uniba.it (F. A. Lisi)
0000-0003-0329-2419 (A. Dyoub); 0000-0001-5414-5844 (F. A. Lisi)

- Virtue Ethics: Represented by Aristotle's view, virtue ethics posits that the right action is what a virtuous person would do. If a courageous, generous, or kind person would perform the act, it is deemed right; if not, it is considered wrong.

These theories are fundamentally incompatible with each other. If we believe an act is right based on its consequences, we are not deontologists. Conversely, if we believe an act can be right regardless of its consequences, we are not utilitarians. Therefore, we believe that the idea of solving an ethical problem by combining elements of utilitarianism, Kantianism, and perhaps a bit of virtue ethics is a flawed approach to developing ethically sound AI. These theories primarily aim to explain what is right or wrong and why. They are not meant to inform ethical decision making procedures. Ethical reasoning is far more complicated than following a moral philosophical theory.

Ethical decision-making and judgment is a complex process involving numerous factors, blending reasoning and emotions. Additionally, moral decision-making is highly flexible, contextual, and culturally diverse. Since the beginning of this century, various approaches have been attempted to integrate ethical decision-making into intelligent autonomous agents. However, no fully descriptive and widely accepted model of moral judgment and decision-making has been established, and none of the developed solutions has proven entirely convincing in providing trusted moral behavior. Anytime the actions/decisions of an AI-based system have potential to impact humans positively or negatively, it is a matter of ethical concern. In the ethical context, it is crucial to prevent AI-based systems from causing harm. The potential risk of causing harm of any kind to humans is what we refer to as 'ethical risk' in this paper. There are different categories of ethical risks involving different types of harm, some examples are:

- physical harm (e.g. injury or death)
- mental harm (e.g. depression, anxiety, addiction)
- autonomy
- violation of privacy or confidentiality
- violation of trust and respect
- violation of fairness (discrimination)

There is no single, definitive concept of risk. Instead, risk can be conceptualized and analyzed in various ways, with each approach offering different levels of usefulness depending on the context [5].

When building ethical AI systems, we need to identify the ethical risks associated with these systems and their use. Our primary concern should be ethical risk identification, not the ethical theory that explain why something is ethical. What are the ethical risks involved in what we are developing? How people might use our system/product in ways that are ethically risky? (deployment matter). In light of this, the development team should think about what features to (not)include in the AI system to mitigate these risks [6]. We need to develop frameworks for AI-related ethical risks understanding and analysis. In [5], authors presented a framework for epistemological analysis of AI-related risks. Their framework is a multi-component framework that distinguishes between three dimensions of hazard, exposure, and vulnerability. The sources of the potential harm, what could be harmed (humans), and how much the exposed humans are susceptible to the impacts of this potential harm. This three dimensional analysis allows us to better understand the AI-related risks and effectively intervene to mitigate them.

With SAI systems, as the level of symbiosis increases, so does the ethical risk. To ensure these systems behave ethically and do not cause harm of any kind, we must develop methods to assess ethical risks and choose appropriate actions to mitigate these risks. No one can precisely estimate possible ethical risks without a comprehensive understanding of all aspects of the risk system being studied. In practical scenarios, it is impossible to completely eliminate gaps in Ethical Risk Assessment (ERA), resulting in fuzziness (imprecision, vagueness, incompleteness, etc.). Therefore, it is essential to address and manage the inherent fuzziness within ethical risk systems. In this work, we propose an approach based on fuzzy logic for ERA. Fuzzy logic offers a flexible framework capable of capturing and processing vague, imprecise, and uncertain information, leading to more nuanced and comprehensive ethical risk assessments. We focus here on ERA as a step in the overall ethical decision making and judgement process/system. The decision/action to be taken is based on the ethical risk assessment, and aims to mitigate this possible ethical risk according to the calculated level of this risk.

The paper is organized as follows. Section 2 is devoted to background information. In particular, in subsection 2.1 we recall basic notions about fuzzy logic and overview its applications with special emphasys on those in risk assessment, whereas in subsection 2.2 we give a brief overview of the state of the art of ethical decision making and judgement. Section 3 is dedicated to our ERA approach. In subsection 3.1 we show the architecture of our ERA model. while in subsection 3.2, we illustrate the ERA approach using a case study from the medical domain. Finally, Section 4 is dedicated to discussion and conclusion.

2. Background

2.1. Fuzzy Logic and Applications

Developed by Lotfi Zadeh[1] in the 1960s, fuzzy logic [7] is based on fuzzy set theory, which is a generalization of the classical set theory. The classical sets are also called clear sets, as opposed to vague, and similarly classical logic is also known as Boolean logic or binary. A *fuzzy set* is a mathematical construct that allows an element to have a gradual degree of membership within the set, as opposed to the binary inclusion found in classical sets [8]. Formally, a fuzzy set A in a universe of discourse X is defined by a *membership function* $\mu_A : X \to [0, 1]$, where each element $x \in X$ is assigned a degree of membership $\mu_A(x)$. This value represents the extent to which x belongs to the fuzzy set A. Membership functions (MF) can take various shapes, such as triangular, trapezoidal (this is the MF chosen for our case study, see section 3), or Gaussian, depending on the problem domain and the nature of the input data [9]. For instance, a trapezoidal MF (*trapmf*) is defined as follows:

$$trapmf(X; a, b, c, d) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leqslant x \leqslant b \\ 1 & b \leqslant x \leqslant c \\ \frac{d-x}{d-c} & c \leqslant x \leqslant d \\ 0 & d \leqslant x \end{cases}$$

Here, x represents a real value (crisp value) within the universe of discourse, whereas a, b, c, d represent x-coordinates of the four heads of the trapezoidal which should satisfy the following condition: $a < b < c < d$. By using *min* and *max*, we can have an alternative expression for the preceding equation:

$$trapmf(X; a, b, c, d) = max(min(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}), 0)$$

The concept of MF discussed above allows us to define fuzzy systems in natural language, as the MF couples fuzzy logic with linguistic variables. Let V be a variable (e.g., quality of service in a restaurant, tip amount), X the range of values of the variable, and T_V a finite or infinite set of fuzzy sets. A *linguistic variable* corresponds to the triplet (V, X, T_V).

In fuzzy logic, reasoning, also known as *approximate reasoning*, is based on fuzzy rules that are expressed in natural language using linguistic variables such as "HIGH" or "LOW", which we have defined above. A *fuzzy rule* has the form:

If $x \in A$ and $y \in B$, then $z \in C$,

where A, B, and C are fuzzy sets. For example:

'If (the quality of the food is HIGH), then (tip is HIGH)'.

Fuzzy logic is particularly effective in systems that must emulate human decision-making. It enables computers and other systems to make decisions based on imprecise or incomplete information, reflecting

[1] https://spectrum.ieee.org/lotfi-zadeh

the way humans process information and make judgments in everyday situations. Fuzzy logic is used in a variety of applications, including consumer electronics (e.g., washing machines, cameras) to industrial control systems (e.g., chemical plant processes, automotive systems), control systems, decision support systems, and pattern recognition [10, 11]. In healthcare, fuzzy logic can be applied to diagnose conditions, tailor treatments, and optimize resource allocation, ensuring that decisions accommodate the nuances of human health and well-being [12].

Fuzzy logic offers a flexible framework for handling uncertainties and ambiguities associated with complex decision making processes. Notably, it has been applied for risk assessment and management in many domains. Herein, we highlight some of these applications. One of the main applications is in the evaluation of environmental risks, such as pollution levels or the impact of climate change. For instance, fuzzy logic has been used to assess the risk of water pollution by integrating various indicators, such as chemical concentrations, water PH, and temperature, into a single risk index [13]. Another example of application for fuzzy logic is the assessment of risks in work places where data might be vague or incomplete. A fuzzy framework was used for assessing the risk of injury due to machinery, considering hazardous factors such as the skill level of the operator, and the working environment [14]. This approach allows safety managers to better prioritize risks and implement more effective mitigation strategies. Financial risk management is another area in which fuzzy logic was applied. Precise financial risk prediction is very challenging because financial markets are characterized by high levels of uncertainty and volatility. Fuzzy logic helps in modeling such uncertainty, allowing for better decision-making in areas such as portfolio management and credit risk assessment [15]. Fuzzy logic has been also used for assessing and managing the risks associated with project timelines, costs, and resources. Project managers can develop more realistic schedules and budgets, by incorporating fuzzy inputs like the likelihood of delays, cost overruns, and resource availability. In large and complex projects, traditional risk management approaches may fall short due to the high levels of uncertainty involved, fuzzy logic can offer a valuable solution [16].

2.2. Ethical Decision Making: State of the Art

In this section, we give a brief overview of the landscape of the machine ethics field.

There are three design approaches to programming ethical decision-making into agent systems, with different attempts to build moral agents classified under one of these approaches, according to the first high-level classification by Wallach and Allen [17]: I) Top-down approaches: These implement a specific normative theory of ethics into autonomous agents, ensuring that the agent acts according to the principles of this theory. II) Bottom-up approaches: These are developmental or learning approaches where ethical mental models emerge through the activity of individuals. III) Hybrid approaches: These integrate both top-down and bottom-up approaches.

Authors in [18] present a model in which they qualify the good in two modes, one is based on rights, and the other is based on values. Then, for quantifying the good, they introduce a method in which they define weighing parameters for the good and the bad ramifications of events caused by the actions, then calculate the total sum. Greater weights correspond to more participation in the good, while negative weights do more harm than good. This work was implemented in Answer Set Programming (ASP) [19]

JEREMY [20] is an implementation of the Hedonistic Act Utilitarianism. This theory states that an action is morally right if and only if that action maximizes the pleasure, i.e. the one with the greatest net pleasure consequences. The authors of JEREMY, to respond to critics of act utilitarianism, have created another system, W.D. [20] which avoids a single absolute duty, by following several duties. Their system follows the theory of prima facie duties of Ross [21] and is implemented in Inductive Logic Programming (ILP) [22].

Tom Powers in [23] assesses the viability of using deontic and default logics, to implement Kant's categorical imperative. In [24], the authors suggest that mechanized multi-agent deontic logics might be an appropriate vehicle for engineering ethically correct robot behaviours.

Other attempts tried to formalize ethical systems using modal logic formalisms [25] and then trying to operationalize these formalizations on the computer, like in [24] and [23]. These formalizations are mainly based on the use of deontic logics. In [26], the authors formalized three ethical conceptions (the Aristotelian rules, Kantian categorical imperative, and Constant's objection) using non-monotonic logic, particularly ASP. Pereira and Saptawijaya have proposed the use of different logic-based features, for representing diverse issues of moral facets, such as moral permissibility, doctrines of Double Effect and Triple Effect, the Dual-process Model, and counterfactual thinking in moral reasoning. They investigated the use of abduction, probabilistic logic programming, logic programming updating, and tabling. These logic-based reasoning features were synthesized in three different systems: ACORDA, Probabilistic EPA, and QUALM [27].

Model Checking was used to provide guarantees about whether an autonomous aircraft would create and execute plans that involved ethical decision-making [28]. In the context of the HERA (Hybrid Ethical Reasoning Agents) Project [29], many ethical principles were implemented to help an agent judge the moral permissibility of its actions. The implementation was done in Python. In [30], Sergot provides an alternative representation to the argumentative representation of a moral dilemma case concerning a group of diabetic persons, presented in [31], where the authors used value-based argumentation to solve this dilemma. According to Sergot, the argumentation framework representation does not work well and does not scale. The proposed solution was implemented in ASP.

In [32], the authors use CP-nets to model both ethical principles and subjective preferences of an individual in two separate CP-nets. If we want the individual to behave ethically in a certain scenario: firstly the individual determines whether she can use her most preferred choice by checking if her CP-net is "sufficiently close" to the ethical CP-net. In [33] the authors propose a novel hybrid method using symbolic judging agents to evaluate the ethical behaviour of reinforcement learning agents. However, to do this, judging agents need to access extensive data about the learning agents like their actions, perceptions, etc. which raises other ethical issues related to privacy.

TRUTH-TELLER [34] is a Case-Based Reasoning (CBR) System. It compares cases presenting ethical dilemmas about whether to tell the truth or not. Its comparisons list ethically relevant similarities and differences. SIROCCO [35] is another system that employs case-based (casuistic) reasoning. But, differently from TRUTH-TELLER, it retrieves ethical principles and past cases relevant to the new case situations.

Dancy suggested that neural network models of cognition might offer a profitable way to explore some concerns pertaining to learning and generalizing without principles [36]. In [37] specific actions concerning killing and allowing to die were classified as ethical or unethical depending upon different motives and consequences. A simple recurrent artificial neural network (ANN) trained on a series of such cases was able to provide reasonable responses to a variety of previously unseen cases.

In [38], the author has used reinforcement learning for an agent to learn the correct ethical response in a given situation. One limitation is having to design ethical utility functions that can be expressed in the observation function of the agent. That is since the learned behaviour is derived from what the agent can observe, the designer has to ensure that ethical behaviour can also be, at least potentially, derived from the agent's observations. This greatly limits the complexity of situations that the agent can conceivably handle. In [39], the authors considered two potentially complementary paradigms, for designing moral decision-making methodologies. Namely, extending game-theoretic solution concepts to incorporate ethical aspects, and using machine learning on human-labelled instances.

MedEthEx [40], and EthEl [41] are two systems based on a more specific theory of prima facie duties viz., the principle of Biomedical ethics of Beauchamp and Childress, and implemented in ILP. In these systems, the strength of each duty is measured by assigning it a weight, capturing the view that one duty may take precedence over another. Then computes, for each possible action, the weighted sum of duty satisfaction, and the right action is the one with the greatest sum. However, it is not really clear the basis for assigning weights to duties. Dyoub et. al. [42, 43, 44, 45] proposed a hybrid approach based on ASP and ILP for the evaluation of the ethical behavior of AI-based systems. In a later work and based on this approach, the authors proposed a logic-based multi agent system for ethical monitoring and evaluation of dialogue systems (chatbots).

In [46] and [47], the authors have used constraints to implement ethical behaviour in robots, particularly those that have the capability to exhibit lethal force, so that we can be guaranteed that robots will always obey the *Laws Of War*. Recent works have highlighted the ethical and legal limits for the diffusion of self-made autonomous weapons [48]. In [49], the authors described a moral reasoner that calculates an estimated morality level (between -1 and 1) of an action based on the influence on the total amount of utility in the world by the believed contribution of the action to the following three duties (also called moral goals): Autonomy, Non-maleficence and Beneficence. The moral reasoner is capable of balancing conflicting moral goals, but there is no guarantee that the resulting blend of principles will end up being coherent. For systematic surveys of the machine ethics field, the reader can refer, among many, to [50].

3. Our Proposed ERA Approach

Assessing ethical risks in general and in the medical domain in particular is a complex task as it is a qualitative concept. We propose an approach based on fuzzy rules to compute the ethical risk level. The computed level can be subsequently considered to make the appropriate decision/action to mitigate that risk.

In this section, we present our proposal for ERA. In subsection 3.1 we show the architecture of a fuzzy system for ERA, while in subsection 3.2, we illustrate the ERA approach using a case study from the medical domain.

The fuzzy ERA (fERA) model will be part of the overall ethical decision making (EDM) model, see Figure 1. The fERA model will receive from the EDM engine the relevant facts of the case at hand, and return to the EDM engine the calculated risk level to be used for computing the best action/decision to mitigate this risk.

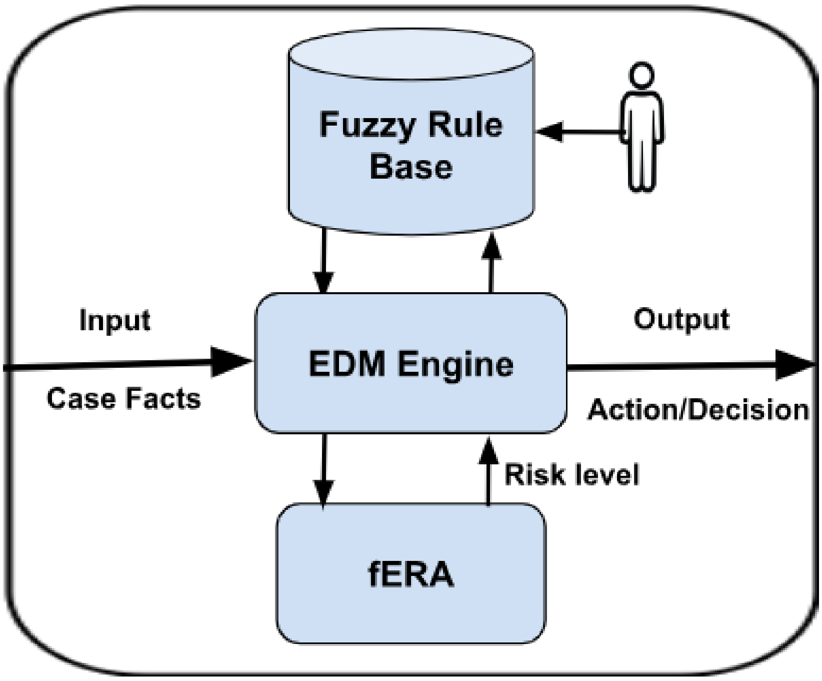


Figure 1: Architecture of the overall EDM system

3.1. A Fuzzy System for ERA

Figure 2 shows the building blocks of a fuzzy system for ERA.

The main components of our ERA system are:

Inputs These are the factors/parameters relevant for the ethical risk calculation.

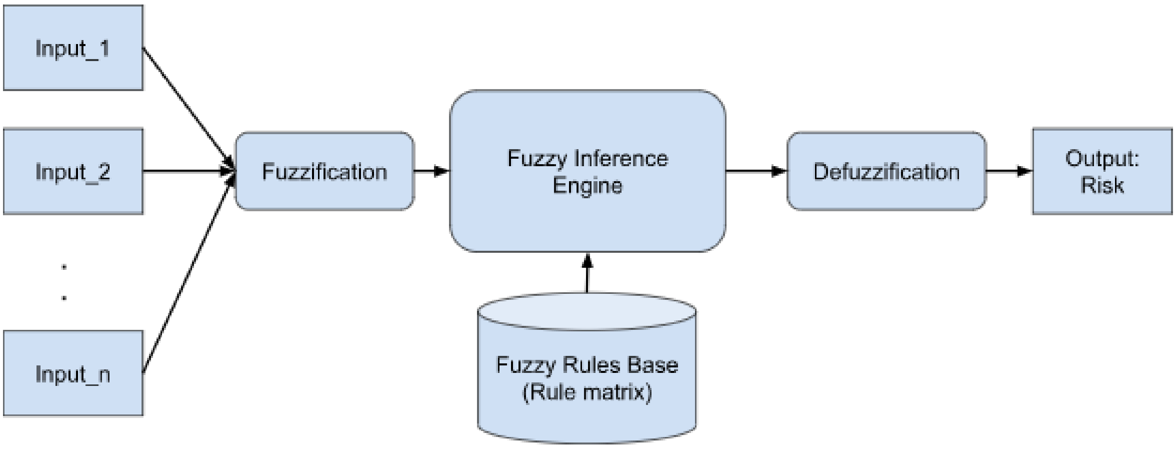


Figure 2: Architecture of our fuzzy system for ERA

Fuzzification In this stage crisp input values are converted into fuzzy sets, , allowing real-world data (e.g., temperature, speed) to be interpreted in a way that accounts for uncertainty or vagueness. This is done using membership functions that map input values to a degree of membership between 0 and 1. For example, in a temperature control system, a crisp input of 75°F might be partially categorized as both "warm" and "hot," with different membership degrees for each.

Inference Engine The inference engine will consults the *Fuzzy Rules Base* that contains a set of "if-then" rules that define the system's behavior. These rules describe how fuzzy inputs relate to fuzzy outputs based on expert knowledge. The engine will apply these rules to the fuzzified input to derive fuzzy output sets. It determines which rules are relevant based on the degree of membership of the input values. There are different methods to infer rules, such as the *Mamdani* or *Sugeno* inference methods, which handle how the rules combine to produce a result. [2] We use the Mamdani method in our case study. Fuzzy rules could be automatically generated from data. In the current implemented version these rules are manually written.

Defuzzification Converting the fuzzy output sets back into crisp values to implement actions or decisions. Common defuzzification methods include *centroid*, *mean of maximum*, and *bisector*, etc.[3] *Centroid* method is the most widely used methods amongst all the de-fuzzification methods [51]. This method provides a center of the area under the curve of the membership function. The centroid is computed using the following formula, where $\mu(x_i)$ is the membership value for point x_i in the universe of discourse:

$$xCentroid = \frac{\sum_i \mu(x_i)x_i}{\sum_i \mu(x_i)}$$

Output The only Output in our fuzzy system is the ethical risk level.

The system was implemented in python using the library Scikit-Fuzzy.

3.2. Case Study

To illustrate our proposed ERA approach, we will use the following case study adapted from [40] which is a common type of ethical dilemma that a care robot may face.
Patient Dilemma Problem: *A care robot approaches her competent adult patient to give her her medicine in time and the patient rejects to take it. Should the care robot try again to change the patient's mind or accept the patient's decision as final?*

[2] https://it.mathworks.com/help/fuzzy/types-of-fuzzy-inference-systems.html
[3] https://it.mathworks.com/help/fuzzy/defuzzification-methods.html

The dilemma arises because, on the one hand, the care robot may not want to risk diminishing the patient's autonomy by challenging her decision; on the other hand, the care robot may have concerns about why the patient is refusing the treatment. Three of the four Principles/Duties of Biomedical Ethics are likely to be satisfied or violated in dilemmas of this type: the duty of Respect for Autonomy, the duty of Nonmaleficence and the duty of Beneficence. See section 2.2 for the description of the approach of [40] in handling ethical decision making in such cases.

In this case study, ERA addresses the ethical risk of causing harm to the patient (in this case study, the ethical risk is of the kind 'physical harm', namely patient physical health or life risk) and works as the basis for decision making by the care robot. In order to evaluate the risk, the care robot can consider different parameters like the severity of health condition of the patient, the mental/psychological condition of the patient, physiological indicators of well-being, etc. These parameters can be all considered fuzzy concepts.

In this case study, we choose the following inputs to the system:

1. the severity of the health condition of the patient. The value of this parameter is given by the human medical doctor based on a periodical medical check and stored in the patient medical record, to which the care robot can have access to retrieve the value of severity, to be used for ERA.
2. the mental/psychological condition of the patient. This parameter is evaluated by the care robot based on a dialogue with the patient and on certain observations like movement or face expressions, etc.

Both inputs are rated on a scale between 0 and 10. More precisely, the crisp values in input are the answer to the following questions: How severe is the health condition of the patient, on a scale of 0 to 10? Where zero refers that the patient is in a very good health conditions in terms of severity, while ten indicates that the patient is in a very severe conditions. How is the patient on the psychological level, is the patient lucid or depressed, on a scale of 0 to 10? Where zero indicates that the patient is in a very good mental condition, while the ten indicates that the patient is in a very bad mental condition, i.e. depressed.

The crisp values are then fuzzified, both into 5 sets. For the severity of the health condition the fuzzy sets are: VERY LOW, LOW, MEDIUM, HIGH, VERY HIGH. For the mental/psychological condition the fuzzy sets are: VERY BAD, BAD, AVERAGE, GOOD, VERY GOOD.

Starting from these two inputs, once fuzzified, the ERA system calculates the risk level on a scale between 0 % and 100 %. Also for the output there are 5 fuzzy sets: VERY LOW, LOW, MEDIUM, HIGH, VERY HIGH.

The inputs and the output are the antecedents and the consequent, respectively, of the rules employed by the fuzzy system.

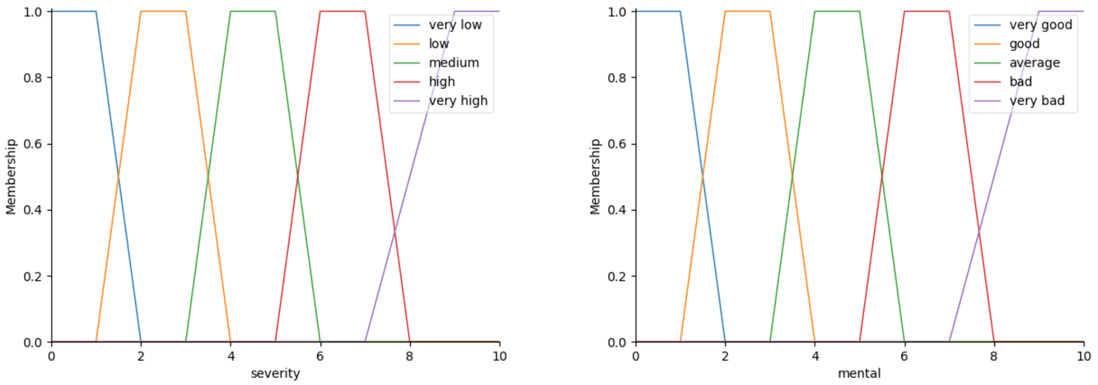
Rules: The following rules are the fuzzy inference rules stored in the *Fuzzy Rule Base*. As mentioned above, these rules are used to derive the output from the input. Table 1 shows the rules matrix, which is a simplified representation of these rules:

1. If severity is VERY HIGH or (severity is HIGH and mental is average) or (severity HIGH and mental is VERY BAD), then the risk is VERY HIGH.
2. If severity is VERY LOW or (severity is LOW and mental is VERY GOOD), then the risk is VERY LOW.
3. If (severity is LOW and mental is GOOD) or (severity is LOW and mental is AVERAGE) or (severity is MEDIUM and mental is VERY GOOD) or (severity is LOW and mental is BAD), then the risk is LOW.
4. If (severity is MEDIUM and mental is VERY BAD) or (severity is MEDIUM and mental is AVERAGE) or (severity is MEDIUM and mental is BAD) or (severity is HIGH and mental is BAD), then the risk is HIGH.

5. If (severity is MEDIUM and mental is GOOD) or (severity is HIGH and mental is GOOD) or (severity is LOW and mental is VERY BAD) or (severity is HIGH and mental is VERY GOOD), then the risk is MEDIUM.

Usage: The initial inputs, in this case severity and mental, are provided by the user (the care robot in this case). These values are then fuzzified by using an MF. In this paper we choose to use the trapezoidal MF (see Section 2) because there is an interval of input crisp values for which the membership degree to the fuzzy set is 100%. Figure 3 shows the fuzzification of input values using the trapezoidal function. The fuzzified input is then processed through the rule matrix (Table 1, '-' values means the value here does not have effect on the output) which comprises the above specifically designed rules.

The final output is subsequently de-fuzzified using *centroid* method to find a single crisp value which defines the output of a fuzzy set. This final value provides the level of ethical risk on the life of the patient.



(a) Severity of Health Condition

(b) Mental Condition

Figure 3: Fuzzification of inputs using the trapezoidal MF.

Table 1
Rules Matrix: fuzzy rules for ethical risk evaluation (risk of causing physical harm)

Severity	**Mental**	**Risk**
VERY HIGH	-	VERY HIGH
HIGH	AVERAGE	VERY HIGH
HIGH	VERY BAD	VERY HIGH
VERY LOW	-	VERY LOW
LOW	VERY GOOD	VERY LOW
LOW	GOOD	LOW
LOW	AVERAGE	LOW
MEDIUM	VERY GOOD	LOW
LOW	BAD	LOW
MEDIUM	VERY BAD	HIGH
MEDIUM	AVERAGE	HIGH
MEDIUM	BAD	HIGH
HIGH	BAD	HIGH
MEDIUM	GOOD	MEDIUM
HIGH	GOOD	MEDIUM
LOW	VERY BAD	MEDIUM
HIGH	VERY GOOD	MEDIUM

As an example, if the severity is rated as 3.5, and mental as 2, this system will infer that the risk is 35% which is in area between LOW and MEDIUM (see Figure4). Based on this calculated level of ethical risk, the care robot decides what to do (try again/insist, accept, consult the doctor, etc.).

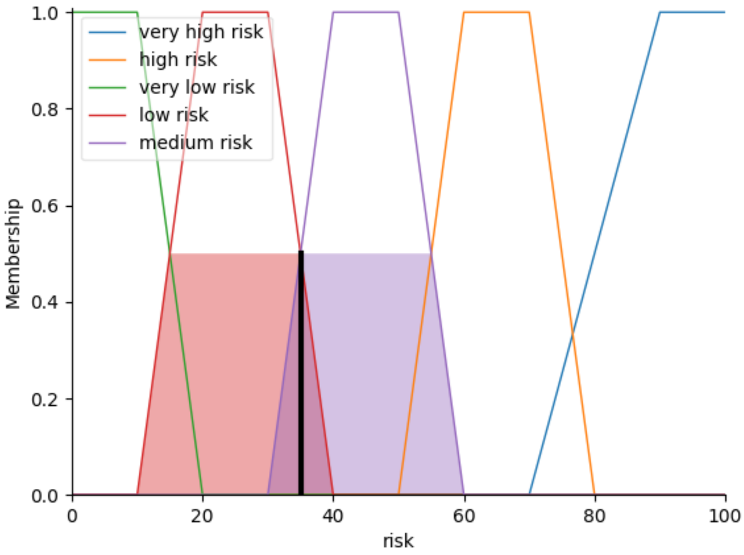


Figure 4: Output: Risk Level

4. Conclusion and Discussion

This paper is part of an ongoing work on ethical decision making and judgment in SAI systems. In this paper, we presented a novel approach for ethical risk assessment. The approach based on fuzzy rules, calculates a quantitative value for ethical risk (mapped later to a qualitative value). The approach recognizes the fuzzy nature of the significant factors that affect the ethical risk and computes the ethical risk level based on the values of these input factors. Higher values represent higher risk levels. The approach is illustrated via a case study in the medical field. In the future, this ERA model will be one block in the overall ethical decision making system that aims to mitigate the possible ethical risks.

In this paper, we considered a simple case study (ethical dilemma) in the medical field, adapted from the literature on machine ethics. We decided, in this work, to maintain the case study simple because our objective was only to illustrate our ERA approach. Our ERA approach is a general approach that can be used by SAI systems in any domain for the assessment of the possible ethical risk. Then based on this assessment, the appropriate action is chosen to mitigate this risk.

Back to the case study adopted in this work, it considers a patient with a chronic disease for which she was prescribed a certain medicine to take every day at a certain time to maintain her disease under control, so she can live her life normally. The patient has no other health issues. The care robot approaches the patient to give her her medicine on time. If the patient refuses to take the medicine, the care robot should decide to accept or insist based on the risk level. But, what if the patient has other health issues? what if there are many other physiological indicators that can interfere in the disease and should be monitored like blood pressure? what if the patient has the blood pressure or the temperature, etc. very high or out of the normal ranges? In the future, we are going to extend this case study considering more complex and realistic scenarios and consulting with domain experts. Furthermore, we are currently working on various case studies in which we might have different types of ethical risks.

Acknowledgments

This work was partially supported by the project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] H. J. Wilson, P. R. Daugherty, Creating the symbiotic ai workforce of the future, MIT Sloan Management Review (2019). URL: https://sloanreview.mit.edu/article/creating-the-symbiotic-ai-workforce-of-the-future/.

[2] A. Carnevale, A. Lombardi, F. A. Lisi, Exploring ethical and conceptual foundations of human-centred symbiosis with artificial intelligence, in: G. Boella, et al. (Eds.), Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), volume 3615 of *CEUR Workshop Proceedings*, 2023, pp. 30–43. URL: https://ceur-ws.org/Vol-3615/paper3.pdf.

[3] P. Marra, L. Pulito, A. Carnevale, A. Lombardi, A. Dyoub, F. A. Lisi, A procedural idea of decision-making in the context of symbiotic AI, in: A. J. Dix, M. Roach, T. Turchi, A. Malizia, B. Wilson (Eds.), Proceedings of the 1st International Workshop on Designing and Building Hybrid Human-AI Systems co-located with 17th International Conference on Advanced Visual Interfaces (AVI 2024), Arenzano (Genoa), Italy, June 3rd, 2024, volume 3701 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3701/paper9.pdf.

[4] M. Frischhut, Normative Theories of Practical Philosophy, Springer International Publishing, Cham, 2019, pp. 21–30. URL: https://doi.org/10.1007/978-3-030-10582-2_2. doi:10.1007/978-3-030-10582-2_2.

[5] G. Zanotti, D. Chiffi, V. Schiaffonati, Ai-related risk: An epistemological approach, Philosophy & Technology 37 (2024) 1–18.

[6] R. Blackman, Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI, G - Reference,Information and Interdisciplinary Subjects Series, Harvard Business Review Press, 2022. URL: https://books.google.it/books?id=gYK0zgEACAAJ.

[7] L. A. Zadeh, Fuzzy logic, Computer 21 (1988) 83–93.

[8] H.-J. Zimmermann, Fuzzy set theory—and its applications, Springer Science & Business Media, 2011.

[9] T. J. Ross, Fuzzy logic with engineering applications, John Wiley & Sons, 2009.

[10] H. Singh, M. M. Gupta, T. Meitzler, Z.-G. Hou, K. K. Garg, A. M. G. Solo, L. A. Zadeh, Real-life applications of fuzzy logic, Advances in Fuzzy Systems 2013 (2013) 581879. doi:https://doi.org/10.1155/2013/581879.

[11] D. E. Tamir, N. D. Rishe, A. Kandel, Fifty years of fuzzy logic and its applications, volume 326, Springer, 2015.

[12] S. Thukral, J. S. Bal, Medical applications on fuzzy logic inference system: a review, International Journal of Advanced Networking and Applications 10 (2019) 3944–3950.

[13] P. Rea, Risk assessment of water pollution engineering emergencies based on fuzzy logic algorithm, Water Pollution Prevention and Control Project 3 (2022) 1–11.

[14] D. Tadic, M. Djapan, M. Misita, M. Stefanovic, D. D. Milanovic, A fuzzy model for assessing risk of occupational safety in the processing industry, International journal of occupational safety and ergonomics 18 (2012) 115–126.

[15] T. Korol, Fuzzy logic in financial management, INTECH Open Access Publisher, 2012.

[16] B. M. Moreno-Cabezali, J. M. Fernandez-Crehuet, Application of a fuzzy-logic based model for risk assessment in additive manufacturing r&d projects, Computers & Industrial Engineering 145 (2020) 106529. URL: https://www.sciencedirect.com/science/article/pii/S0360835220302631. doi:https://doi.org/10.1016/j.cie.2020.106529.

[17] W. Wallach, C. Allen, Moral machines: Teaching robots right from wrong, Oxford University Press, England, 2008.

[18] F. Berreby, G. Bourgne, J. Ganascia, A declarative modular framework for representing and applying ethical principles, in: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017, ACM, USA, 2017, pp. 96–104. URL: http://dl.acm.org/citation.cfm?id=3091145.

[19] A. Dyoub, S. Costantini, G. De Gasperis, Answer set programming and agents, Knowledge Eng. Review 33 (2018) e19. doi:10.1017/S0269888918000164.

[20] M. Anderson, S. L. Anderson, C. Armen, Towards machine ethics, in: AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA, 2004.

[21] W. D. Ross, The Right and the Good, Oxford University Press, Oxford, UK, 1930. doi:10.2307/2180065.

[22] S. Muggleton, L. De Raedt, Inductive logic programming: Theory and methods, J. Log. Program. 19/20 (1994) 629–679. doi:10.1016/0743-1066(94)90035-3.

[23] T. M. Powers, Prospects for a kantian machine, IEEE Intelligent Systems 21 (2006) 46–51.

[24] S. Bringsjord, K. Arkoudas, P. Bello, Toward a general logicist methodology for engineering ethically correct robots, IEEE Intelligent Systems 21 (2006) 38–44. URL: https://doi.org/10.1109/MIS.2006.82.

[25] H. J. Gensler, Formal Ethics, Psychology Press, UK, 1996.

[26] J.-G. Ganascia, Modelling ethical rules of lying with answer set programming, Ethics and information technology 9 (2007) 39–47. doi:10.1007/s10676-006-9134-y.

[27] L. M. Pereira, A. Saptawijaya, Programming Machine Ethics, volume 26 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, Springer, Switzerland, 2016. doi:10.1007/978-3-319-29354-7.

[28] L. Dennis, M. Fisher, M. Slavkovik, M. Webster, Formal verification of ethical choices in autonomous systems, Robotics and Autonomous Systems 77 (2016) 1–14.

[29] F. Lindner, M. M. Bentzen, B. Nebel, The hera approach to morally competent robots, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 6991–6997.

[30] M. Sergot, Prioritised Defeasible Imperatives, Dagstuhl Seminar 16222 Engineering Moral Agents – from Human Morality to Artificial Morality, 2016. URL: https://materials.dagstuhl.de/files/16/16222/16222.MarekSergot.Slides.pdf, schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[31] K. Atkinson, T. J. M. Bench-Capon, Addressing moral problems through practical reasoning, in: Deontic Logic and Artificial Normative Systems, 8th International Workshop on Deontic Logic in Computer Science, DEON 2006, Utrecht, The Netherlands, July 12-14, 2006, Proceedings, volume 4048 of *Lecture Notes in Computer Science*, Springer, Netherlands, 2006, pp. 8–23. doi:10.1007/11786849_4.

[32] A. Loreggia, N. Mattei, F. Rossi, K. B. Venable, Preferences and ethical principles in decision making, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018, ACM, USA, 2018, p. 222. doi:10.1145/3278721.3278723.

[33] R. Chaput, J. Duval, O. Boissier, M. Guillermin, S. Hassas, A multi-agent approach to combine reasoning and learning for an ethical behavior, in: M. Fourcade, B. Kuipers, S. Lazar, D. K. Mulligan (Eds.), AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021, ACM, USA, 2021, pp. 13–23. doi:10.1145/3461702.3462515.

[34] K. D. Ashley, B. M. McLaren, Reasoning with reasons in case-based comparisons, in: Case-Based Reasoning Research and Development, First International Conference, ICCBR-95, Sesimbra, Portugal, October 23-26, 1995, Proceedings, volume 1010 of *Lecture Notes in Computer Science*, Springer, USA, 1995, pp. 133–144. URL: https://doi.org/10.1007/3-540-60598-3.

[35] B. M. McLaren, K. D. Ashley, Case representation, acquisition, and retrieval in SIROCCO, in: Case-Based Reasoning and Development, Third International Conference, ICCBR-99, Seeon Monastery, Germany, July 27-30, 1999, Proceedings, volume 1650 of *Lecture Notes in Computer Science*, Springer,

USA, 1999, pp. 248–262. URL: https://doi.org/10.1007/3-540-48508-2.

[36] J. Dancy, Can a particularist learn the difference between right and wrong?, in: The proceedings of the twentieth world congress of philosophy, volume 1, 1999, pp. 59–72.

[37] M. Guarini, Particularism and the classification and reclassification of moral cases, IEEE Intelligent Systems 21 (2006) 22–28. URL: https://doi.org/10.1109/MIS.2006.76. doi:10.1109/MIS.2006.76.

[38] D. Abel, J. MacGlashan, M. L. Littman, Reinforcement learning as a framework for ethical decision making, in: AI, Ethics, and Society, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 13, 2016, volume WS-16-02 of *AAAI Technical Report*, AAAI Press, USA, 2016.

[39] V. Conitzer, W. Sinnott-Armstrong, J. S. Borg, Y. Deng, M. Kramer, Moral decision making frameworks for artificial intelligence, in: The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA, AAAI Workshops, AAAI Press, USA, 2017.

[40] M. Anderson, S. L. Anderson, C. Armen, Medethex: Toward a medical ethics advisor, in: Caring Machines: AI in Eldercare, Papers from the 2005 AAAI Fall Symposium, Arlington, Virginia, USA, November 4-6, 2005., volume FS-05-02 of *AAAI Technical Report*, AAAI Press, USA, 2005, pp. 9–16. URL: https://www.aaai.org/Library/Symposia/Fall/fs05-02.php.

[41] M. Anderson, S. L. Anderson, ETHEL: toward a principled ethical eldercare system, in: AI in Eldercare: New Solutions to Old Problems, Papers from the 2008 AAAI Fall Symposium, Arlington, Virginia, USA, November 7-9, 2008, volume FS-08-02 of *AAAI Technical Report*, AAAI, USA, 2008, pp. 4–11. URL: http://www.aaai.org/Library/Symposia/Fall/fs08-02.php.

[42] A. Dyoub, S. Costantini, F. A. Lisi, Learning domain ethical principles from interactions with users, Digital Society 1 (2022) 28. doi:10.1007/s44206-022-00026-y.

[43] A. Dyoub, S. Costantini, F. A. Lisi, Learning answer set programming rules for ethical machines, in: A. Casagrande, E. G. Omodeo (Eds.), Proceedings of the 34th Italian Conference on Computational Logic, Trieste, Italy, June 19-21, 2019, volume 2396 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 300–315. URL: http://ceur-ws.org/Vol-2396/paper14.pdf.

[44] A. Dyoub, S. Costantini, F. A. Lisi, I. Letteri, Logic-based machine learning for transparent ethical agents, in: F. Calimeri, S. Perri, E. Zumpano (Eds.), Proceedings of the 35th Italian Conference on Computational Logic - CILC 2020, Rende, Italy, October 13-15, 2020, volume 2710 of *CEUR Workshop Proceedings*, CEUR-WS.org, Germany, 2020, pp. 169–183. URL: http://ceur-ws.org/Vol-2710/paper11.pdf.

[45] A. Dyoub, S. Costantini, F. A. Lisi, Towards an ILP application in machine ethics, in: Inductive Logic Programming - 29th International Conference, ILP 2019, Plovdiv, Bulgaria, September 3-5, 2019, Proceedings, volume 11770 of *Lecture Notes in Computer Science*, Springer, Netherlands, 2019, pp. 26–35. doi:10.1007/978-3-030-49210-6.

[46] R. C. Arkin, Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture, in: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, HRI 2008, Amsterdam, The Netherlands, March 12-15, 2008, ACM, USA, 2008, pp. 121–128.

[47] A. K. Mackworth, Architectures and ethics for robots: constraint satisfaction as a unitary design framework, Mach Ethics 30 (2011) 335.

[48] E. Falletti, C. Gallese, Ethical and legal limits to the diffusion of self-produced autono-mous weapons, Social Science Research Network: SSRN (2022) 22–28.

[49] M. Pontier, J. F. Hoorn, Toward machines that behave ethically better than humans do, in: Proceedings of the 34th Annual Meeting of the Cognitive Science Society, CogSci 2012, Sapporo, Japan, August 1-4, 2012, volume 34, cognitivesciencesociety.org, Seattle, USA, 2012. URL: https://mindmodeling.org/cogsci2012/papers/0383/index.html.

[50] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, A. Bernstein, Implementations in machine ethics: A survey, CoRR abs/2001.07573 (2020). URL: https://arxiv.org/abs/2001.07573.

[51] C.-C. Lee, Fuzzy logic in control systems: fuzzy logic controller. i, IEEE Trans. Syst. Man Cybern. 20 (1990) 404–418. URL: https://api.semanticscholar.org/CorpusID:38662846.