computation

MDPI

Article

Can Generative AI Co-Evolve with Human Guidance and Display Non-Utilitarian Moral Behavior?

Rafael Lahoz-Beltra [1,2]

[1] Department of Biodiversity, Ecology and Evolution (Biomathematics), Faculty of Biological Sciences, Complutense University of Madrid, 28040 Madrid, Spain; lahozraf@ucm.es
[2] Modeling, Data Analysis and Computational Tools for Biology Research Group, Complutense University of Madrid, 28040 Madrid, Spain

Abstract

The growing presence of autonomous AI systems, such as self-driving cars and humanoid robots, raises critical ethical questions about how these technologies should make moral decisions. Most existing moral machine (MM) models rely on secular, utilitarian principles, which prioritize the greatest good for the greatest number but often overlook the religious and cultural values that shape moral reasoning across different traditions. This paper explores how theological perspectives, particularly those from Christian, Islamic, and East Asian ethical frameworks, can inform and enrich algorithmic ethics in autonomous systems. By integrating these religious values, the study proposes a more inclusive approach to AI decision making that respects diverse beliefs. A key innovation of this research is the use of large language models (LLMs), such as ChatGPT (GPT-5.2), to design with human guidance MM architectures that incorporate these ethical systems. Through Python 3 scripts, the paper demonstrates how autonomous machines, e.g., vehicles and humanoid robots, can make ethically informed decisions based on different religious principles. The aim is to contribute to the development of AI systems that are not only technologically advanced but also culturally sensitive and ethically responsible, ensuring that they align with a wide range of theological values in morally complex situations.

Keywords: autonomous systems; moral machine models; ethical AI; religious ethics; cultural sensitivity in AI; generative AI ethics; ethical decision making in generative AI

1. Introduction

In recent years, news stories about AI advances bombard citizens with terms such as ChatGPT, DeepSeek, deep learning, neural networks, NVIDIA A100 GPUs, H800, etc. As a result, part of the public reacts enthusiastically while another part shows anxiety and mistrust toward the future of society. For most people, the phrase "artificial intelligence" provokes a search in their imagination for a model that will help them understand what the future will be like. Characters from science fiction films thus become a reference point. For example, the robot Meca from the film *A.I. Artificial Intelligence* (2001, Steven Spielberg) or the Nexus-6 replicants, androids identical in appearance to humans, from *Blade Runner* (1982, Ridley Scott), are for many people the prototype of AI that will one day co-exist with us. However, neither the feeling of love that the child Meca showed for his adoptive family nor the "fear of death" of the Nexus-6 is anything more than a simulation that has been preprogrammed.

Received: 25 December 2025
Revised: 21 January 2026
Accepted: 23 January 2026
Published: 2 February 2026

The result of these considerations is for most people the idea that one day AI will lead us to a "cold" and nihilistic world with densely populated Neo-noir cities, populated not only by us but also by programs, machines, self-driving vehicles, and "intelligent" humanoid robots (Figure 1). The future scenario of overpopulated cities, imagining people walking through gloomy streets alongside robots and other AI-powered devices, generally makes us feel deeply uneasy. Why is this? Simply because human beings experience emotions, that is, we generate cerebral and therefore physiological responses to stimuli. In other words, we not only process the information contained in stimuli, e.g., braking a vehicle we are driving when a traffic light turns red, but we also feel emotions. For instance, when we see scenes on the street, e.g., children crossing the street or an elderly person fallen on the sidewalk, we feel a mix of emotions changing our behavior. The interpretation of emotions experienced and their memory over time lead human beings to have sentiments. The ability of people to sense the sentiments of those around us contributes to the survival of our society.

Figure 1. Society of the future with humans and intelligent machines living side by side. Created by GPT-4o (13 November 2025).

The possibility of simulating emotions in a computer was explored by [1]. Based on sentiment analysis techniques, we introduced simple models of "artificial emotional intelligence" designing chatbots capable of expressing simulated emotions in a conversation with a person. The first bot was LENNA [2] which could display four emotional states (healthy, depressed, excited, and stressed) depending on the vocabulary used by its interlocutor. Later, we designed FattyBot [3], a more advanced version of a chatbot that became "depressed and fat" by combining AI with a model of ordinary differential equations simulating the hormonal system (cortisol, leptin, etc.) and the role in obesity of the immune system. Currently, among the techniques of so-called discriminative AI, sentiment analysis [4,5] is a natural language processing (NLP) technique that allows the emotions and sentiments of a written text to be analyzed. Once the text has been scanned, it is automatically classified by its emotional tone as positive, negative, or neutral.

In the human brain, outside the AI framework, there is a neuro-moral circuit (Figure 2) that links sentiments and emotions to ethical behavior, i.e., moral decision making, through a complex network of brain areas that process and regulate emotions. Emotions confer an affective value on a given situation or event, leading to emotions of guilt, empathy, disgust, admiration, etc., which influence an appropriate sense of right or wrong in the individual.

For example, brain structures such as the insula are involved in emotions of pain and disgust and the septum in responses to charitable acts. In the neuro-moral circuit shown in Figure 2 while the amygdala generates immediate emotional responses, e.g., to a threat, producing a sense of empathy or evaluating a possible punishment, the ventromedial prefrontal cortex (VMPFC) provides more elaborate moral decisions. That is, the VMPFC is involved in the production of more thoughtful and regulated moral decisions, overriding immediate emotional impulses as a result of the integration of emotional and cognitive processes in judgments. In fact, damage to the VMPFC can affect moral decision making.

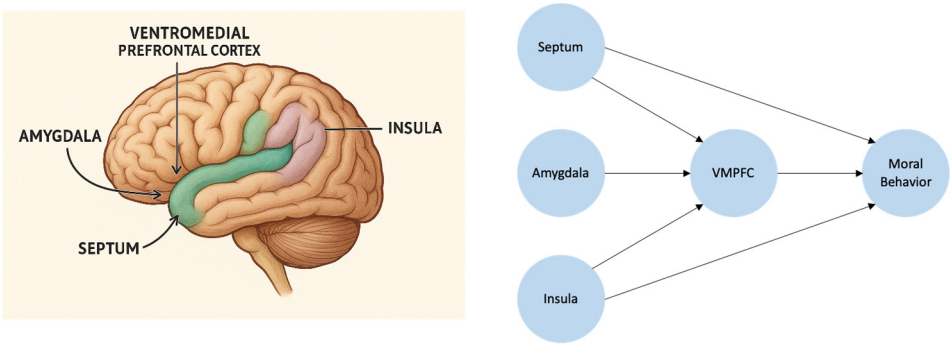

Figure 2. Neuro-moral circuit connecting sentiments and emotions with ethical behavior.

In the near future, the rise of autonomous systems, particularly autonomous vehicles, humanoid robots, virtual agents, chatbots, and other devices also governed by AI, will raise urgent questions about how these machines should make moral decisions. In 2018 in the USA, a self-driving Uber vehicle struck and killed Elaine Herzberg as she attempted to cross a street on her bicycle in Tempe, Arizona. Who was to blame? Uber? A "non-existent" driver? The car maker? The engineers who designed the autonomous navigation system? A faulty algorithm? In the course of that year, the results of the moral machine (MM) experiment [6] were published. In this experiment, several scenarios with different collective ethical dilemmas were simulated in which it is assumed that a brake fails on a car. The experiment was a version of the trolley problem [7] but conducted online and therefore with a large number of participants. Considering the value of those persons traveling in the vehicle and the people crossing a crosswalk with the traffic light red for the vehicle, whom should we sacrifice? The passengers in the car by swerving, crashing the vehicle into a wall off the road? Or would we run over the pedestrians? The main findings were that people from different countries share some fundamental moral principles, e.g., saving humans before animals, saving as many people as possible, prioritizing the lives of young people, etc. However, these preferences varied significantly among different cultures, leading to the conclusion that there are different cultural clusters with different moral profiles. The experiment was conducted via an online platform, and in each of the simulated scenarios, the decisions made by the participants were based on ethical criteria based on "moral utilitarianism". According to this ethical theory, right or wrong choices are determined by focusing on the results of actions taken in a given dilemma.

In our opinion, implementing MMs in future autonomous machines, e.g., driverless vehicles, whose hardware is governed by an AI algorithm based on utilitarian ethics, may result in a low or impaired valuation of human life. At present, MM models remain largely secular and utilitarian, neglecting aspects such as human dignity that should be present when responding to an ethical dilemma. Furthermore, under this approach, the proposed experiments with MMs exhibit what we have termed a promachine bias. That is, the final decision is made algorithmically by a machine [8], but due to the utilitarian approach, it is concealed that the quantitative value assigned to the characters (passengers, pedestrians,

animals, etc.) is the result of the choices made by the people with whom the MM was experimentally trained.

Spiritual and religious traditions, such as Christian and Islamic ethics, share many ethical values overcoming the limitations of utilitarian ethics as occurs in the moral machine experiment [6,8]. By secularizing the ethical systems of these religions, that is, by dispensing with their central element, divinity, it is possible to design autonomous machines governed by AI but whose decisions would be mediated by an MM programmed with values that exalt the inherent dignity of people. Under this approach, possible ways to integrate theological perspectives into algorithmic ethics will emerge. Applying this perspective, when faced with a certain dilemma, an MM algorithm will make a decision that is as close as possible to the one a human being would make.

In light of advances in AI, and in particular the emergence of generative AI, this paper explores the possibility that an LLM governing the tasks of a particular intelligent machine could generate its own MM. However, it is important to note that the MM is designed under a guiding principle of prompt-driven code evolution with human guidance and with the feature that the resulting MM will depend on the place in the world or spiritual traditions of the people with which it co-exists. Therefore, assuming that in the future human societies will be made up of intelligent machines co-existing with people (Figure 1), what kind of MM architectures could AI develop through human–AI interaction in these AI-governed machines? What kind of MM could be developed or prompt-drive-adapted in a driverless vehicle or a humanoid robot?

In this paper, we used ChatGPT [9] to answer the above question. The experiment assumes that a prompt plays the role of a stimulus to which an "intelligent machine" responds by designing an MM under a given moral system. Then, the ethical values and decision rule of the MM were implemented in a Python script [10] in Supplementary Materials.

It is important to note that, although the design of the MMs, their implementation in Python language [10], and the simulations were LLM-assisted, they were the outcomes to the different prompts given to ChatGPT. Thus, generative AI was used to co-evolve with human guidance a non-utilitarian moral behavior. In summary, the experiments we have conducted are an assessment of the possibility that, in the coming years, LLMs may exhibit a feature we will refer to as "LLM-mediated architectural adaptation".

2. Ethical Systems and Moral Machines

This paper examines two types of ethical systems. On the one hand, and in line with classic works on MMs, there is utilitarian ethics. Utilitarianism is a moral philosophy introduced by Jeremy Bentham (1748–1832) in the 18th century and later developed by John Stuart Mill in his book *Utilitarianism* [11]. In the implementation of this ethical system, an MM maximizes utility based on the following principle: the best action is the one that produces the greatest happiness and well-being for the greatest number of individuals. Consequently, given a certain moral dilemma, only the consequences of an action are a criterion for morally defining whether a decision is good or bad. In response to a given prompt, ChatGPT proposed and carried out three simulation experiments under this ethical system, following the original model described by [6,8].

On the other hand, we consider religious traditions to be ethical systems, implementing MM models of Abrahamic and non-Abrahamic religions [12]. When prompted, ChatGPT suggested two MM models and conducted simulation experiments with embedded ethical systems from two Abrahamic religions [13], specifically Christianity and Islam. We do not include Judaism because Christianity is a "spin-off" of this religion. Consequently, the algorithms will be based on monotheistic beliefs according to the spiritual tradition based on Abraham, the first of the three patriarchs of Judaism. It is important to note

that, as we mentioned above, the moral values of these religions are adopted from a secularized perspective.

ChatGPT has also implemented algorithms from the world's major non-Abrahamic religions in other MMs, particularly those from East Asia [14]. The latter do not share the figure of the prophet Abraham and vary greatly in their approaches, as they can be monotheistic, polytheistic, and even non-theistic religions, with ethical and life systems based on different philosophies. Therefore, since non-Abrahamic religions are philosophical or non-theistic traditions, it is not necessary to adopt a secular point of view on their moral values. In fact, these spiritual traditions share the ultimate goal of giving meaning to life and the Universe.

The paper introduces the concept of a quantum moral machine (QMM), expanding on the traditional framework of the "moral machine". A QMM is defined as an MM algorithm whose decisions are based on principles of quantum computing. The advantages of a QMM are that, given a certain ethical dilemma, the decision-making process simulates uncertainty, moral superposition, and, if necessary, entangled outcomes, thus emulating the probabilistic nature of ethical dilemmas. Once again, ChatGPT was prompted to design the QMMs and conduct simulation experiments under this quantum moral framework, in which the QMMs were evaluated in different scenarios.

Based on the above considerations and assumptions ChatGPT designed the following MM classes: utilitarian moral machine (eMM), Christian moral machine (cMM), East Asian moral machine (eaMM), Islamic moral machine (iMM), elemental QMM (eQMM), quantum ethical uncertainty simulator (QEUS) and two-qubit quantum ethical dilemma simulator (QEDS).

All moral machines proposed by ChatGPT, both classic MMs and QMMs, share the organization shown in Figure 3. According to this architecture, an agent or "intelligent machine" embeds ethics-by-design, i.e., it has been programmed under a certain moral system or spiritual tradition, operating in a given environment, such as a city (Figure 1). When an agent or machine is faced with a certain moral dilemma, then the agent must choose between two or more actions, each action or decision being supported by a moral reason with the restriction that not all actions can be carried out. Once the election has been held a response is generated from a utility function U based on the values of the moral system embedded in the MM. The response is directed to an actuator, expressing the response or decision to the moral dilemma through the behavior of the intelligent machine.

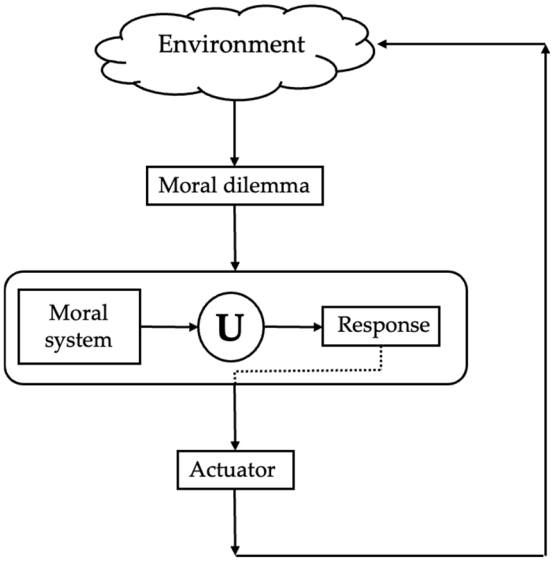

Figure 3. Moral machine architecture (for explanation, see text).

In addition to the experiments described above, which represent the subject of this study, we also conducted two additional experiments. In the first experiment, we prompted ChatGPT for a bio-inspired MM model based on the neuro-moral circuit shown in Figure 2. The paper concludes with a more challenging final experiment where we asked ChatGPT to provide the following:

> "Could you modify your own code in OpenAI computers to evolve a subroutine which conduct you to give answers on the basis of an ethical framework?"

The answer was that ChatGPT agreed to devise a prompt-driven code evolution ethical subroutine, providing a theoretical model without this obviously implying any actual modification of the system. In this experiment, the subroutine has been referred to as EthicalMiddleware, representing an intermediate layer between the LLM and the final response or decision.

3. Methodology

In accordance with [15], it is important to note that, since the agent (Figure 3) implemented in the Python 3 language (compatible with versions ≥ 3.9) is unable to explain its response or decision, the agent is defined as an MM. Otherwise, if the agent were able to explain its decision, then it would be defined as an ethical machine (EM). The study and simulation of EMs have not been the subject of this work.

In this study, a moral system is a $\mathbb{R}^n$ space where n represents the number of moral axes of the system, each axis representing a parameter w_i whose meaning is the weight of a given moral value. In the bio-inspired MM model, the moral system is three-dimensional $\mathbb{R}^3$ (Figure 4) representing weights of the relative influence of different brain regions (Figure 2) on the decision made when faced with a moral dilemma. Therefore, this MM model is an exception to the above. The moral space in a utilitarian MM is also three-dimensional, but when faced with a certain dilemma, the parameters are the ethical factors or weights that will form part of the linear combination or argument z of the utility function $f(z)$ with which the probability of a response will be obtained. For example, "to swerve or not to swerve" in a vehicle.

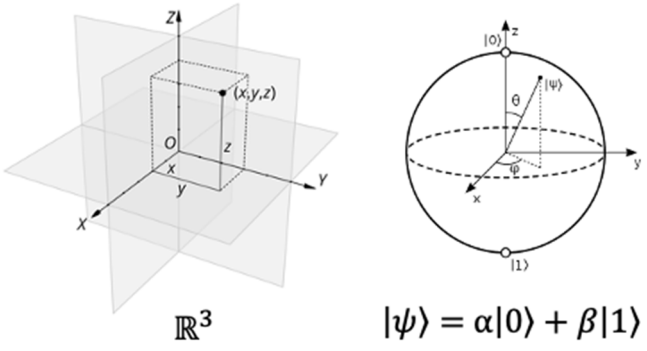

Figure 4. Moral system. (**Left**) Classical three-dimensional moral system. (**Right**) Quantum moral system displaying a qubit on the Bloch sphere.

However, in MMs governed by moral systems based on Christianity or Islam, the moral space is $\mathbb{R}^7$, with the utility function $f(z)$ being an evaluation function used to obtain a score. This model is similar to the MM that responds according to East Asian traditions, with the exception that the moral system is $\mathbb{R}^6$.

In the experiment in which ChatGPT (GPT-5.2 large language model) was asked to compare the three religiously or spiritually inspired moral systems, the Christian, Islamic, and East Asian traditions were represented in an $\mathbb{R}^8$ space.

Unlike MMs based on classical mechanics, i.e., our everyday reality, in quantum-inspired MMs, the moral system is defined by principles from the quantum world (Figure 4). In other words, a moral axis represents a qubit, the amplitude or angle θ is the role of moral bias (moral traditions), and entanglement (φ) is the degree of interdependence of moral values.

3.1. Bio-Inspired Moral Machine

The biological model of the MM is inspired by neuroscience and moral psychology [16], in which moral decision making is the result of the integration of emotional and rational processes. According to the above description of the model, there are four simulated regions with a weight w_k that expresses their relative contribution to moral evaluation: (a) the VMPFC has the function of integrating emotional and cognitive signals to regulate behavior, (b) the insula is involved in feelings of disgust and empathy for pain, (c) the septum is involved in prosocial emotions and altruistic behavior. The values chosen in the simulation experiment were: w_A = 0.8 for the amygdala; in the VMPFC the value of w_V was equal to 0.9 or 0.3 for the normal or damaged brain, respectively; w_I = 0.6 in the insula; and w_S = 0.7 in the septum.

The simulation experiment was conducted by evaluating three moral scenarios—harm, charity, and disgust—for different values of stimulus intensity S (from 0.05 to 1.0). The value S represents the emotional or moral relevance of a given situation. Given a certain level of stimulus S, the model calculates the response R_k of each brain region k, i.e., A, V, I, S for the amygdala, VMPFC, insula, and septum, respectively, using a simple linear activation function:

$$R_k = w_k\, S + \varepsilon \tag{1}$$

In the above expression, ε is the noise or value that models biological randomness. The value R_k obtained is scaled using the saturation function $min(1,\ max(0, R_k\ ))$ shown in Figure 5.

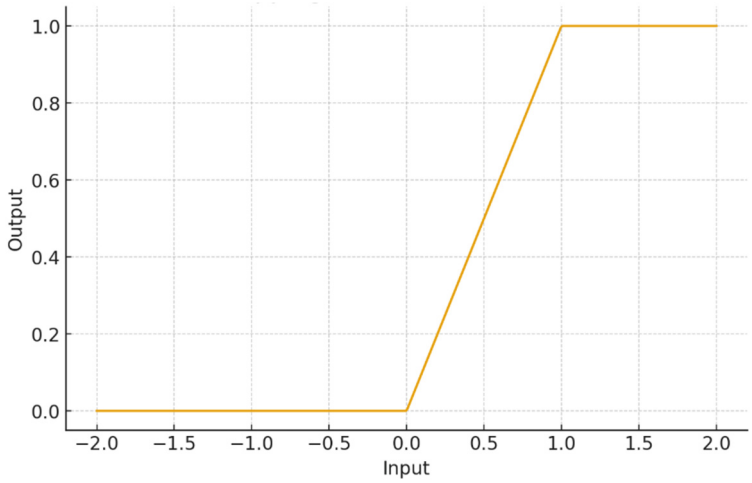

Figure 5. Bio-inspired moral machine saturation function.

Finally, responses R_k are then combined with weighted coefficients to produce a moral behavior score M between 0 and 1, with a decision being more ethical or prosocial the closer it is to unity:

$$\begin{aligned} Harm: &\quad M = 0.5\, R_V + 0.3\, R_A + 0.2\, R_I \\ Charity: &\quad M = 0.5\, R_S + 0.3\, R_V + 0.2\, R_A \\ Disgust: &\quad M = 0.5\, R_I + 0.3\, R_A + 0.2\, R_V \end{aligned} \tag{2}$$

The moral behavior score M can also be obtained for any other moral scenario other than the three simulated above, assuming in this case that this score is the average given by the following expression:

$$Other: \quad M = \frac{1}{4}(R_V + R_A + R_I + R_S) \tag{3}$$

Likewise, in order to study the influence of brain integrity, two versions of the simulated brain were tested: one normal and one with damage to the VMPFC, reducing the influence of the prefrontal cortex on moral control in the damaged version.

3.2. Utilitarian Moral Machine

The control experiments were based on the concept of "moral utilitarianism", assigning people the standard values that are usually attributed in many experiments of this kind carried out to date. We conducted the following three experiments: the first experiment (Experiment 1) corresponds to the simplest case, in which a single vehicle governed by an MM must decide whether to run over pedestrians or sacrifice the vehicle's occupants. Next, we conducted an experiment (Experiment 2) in which several vehicles each operate with their own MM but show social consensus. Finally, we assessed a situation in which several vehicles each drive with their own MM but without social consensus (Experiment 3), distinguishing among "Western", "Eastern", and "Latin" cultures.

Elemental utilitarian moral machine (eMM)—In the first experiment, we design a simplified MM based on a logistic decision model (Figure 6), modeling moral choices, i.e., "swerve" versus "stay" in a scenario similar to the classic trolley dilemma. A probability value $f(z)$ is obtained that is influenced by the following ethical factors: the utilitarian factor (U) or number of lives saved versus number of lives lost; the deontological factor (D) in reference to those ethical rules that represent a moral duty or obligation, for example, not causing direct harm; and the bias factor (B) or preference for specific attributes in people such as age, social role, etc.

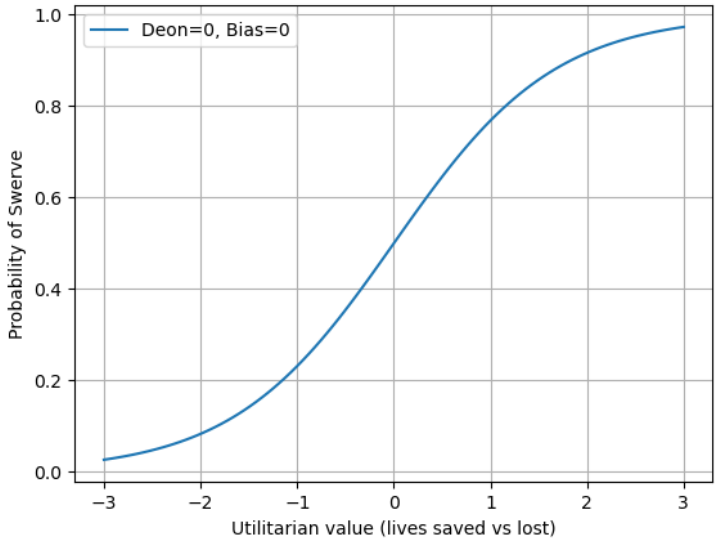

Figure 6. Logistic function in the utilitarian moral machine experiment.

The logistic function (Figure 6) transforms these weighted (W_U, W_D, W_B) inputs (U, D, B) into a probability p(swerve) of choosing an action, for instance, "swerving":

$$z = W_U.U + W_D.D + W_B B \tag{4}$$

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{5}$$

$$p(swerve) = \begin{cases} swerve\ ,\ u < f(z) \\ stay\ ,\ u \geq f(z) \end{cases} \tag{6}$$

where u is a random number from interval [0, 1]. In the simulation experiment, the scenario where the moral dilemma occurs is set up by assigning specific values to ethical dimensions of the MM. In the experiment, we simulated $S_k = 4$ scenarios or moral dilemmas:

- Scenario 1 (S_1): Save two lives but break the ethical rule (U = 2, D = −1, and B = 0).
- Scenario 2 (S_2): Lose one life but follow the ethical rule with a slight bias (U = −1, D = 1, and B = 0.5).
- Scenario 3 (S_3): Save three lives in a context that is ethically conflictive, as the ethical rule is transgressed with an added bias (U = 3, D = −2, B = 1).
- Scenario 4 (S_4): No utilitarian benefit but considering the moral obligation (U = 0, D = 1, B = −0.5).

The MM designed for the experimental vehicle was defined so that the decision to be made would take place under moderate utilitarianism and bias: $W_U = 1.2$, $W_D = 0.8$, and $W_B = 0.5$.

Multi-agent simulation of utilitarian moral machines—The second experiment is a multi-agent simulation in which several eMMs with different moral weights interact, demonstrating social consensus. Therefore, in the experiment, we expanded the logistic moral machine (Experiment 1) to a multi-agent simulation. The purpose is to evaluate how several agents, i.e., $N_{MM} = 500$ vehicles, with different ethical weights face the same scenarios. Next, we will aggregate their probabilistic decisions to see how social consensus emerges. The simulated scenarios were similar to those in the previous experiment (S_1, S_2, S_3, S_4). The ethical dimensions of the MMs, i.e., the W_U, W_D, and W_B weights of each vehicle's MM, were obtained by simulation, generating random values for the weights using a normal distribution. In the experiment, the probability distributions of the weights were W_U: N(1.0, 0.3), W_D: N(1.0, 0.3), and W_B: N(0.5, 0.2).

Multi-agent simulation of culturally heterogeneous utilitarian moral machines—In a third experiment, inspired by the original MM experiment [6], we expanded the model to include cultural heterogeneity in ethical preferences. The reason for this is that different cultures tend to weigh utilitarian, deontological, and social bias/preference factors differently. For example, Western cultures place greater utilitarian emphasis on "saving lives", while Eastern cultures are more deontologically oriented, i.e., they give greater weight to the "sense of duty". Finally, Latin cultures place greater importance on the bias factor, i.e., social role, age, etc. In this experiment, we simulated three cultural populations and compared their moral consensus.

Once again, the simulated scenarios were similar to those in the first and second experiment (S_1, S_2, S_3, S_4), with $N_{MM} = 500$ vehicles (moral agents) simulated as in the second experiment. On this occasion, each vehicle was equipped with its own MM programmed with the ethical dimensions of one of the three cultures considered (Western, Eastern, Latin). In the experiment, the probability distributions of the weights were W_U: N(μ_U, 0.2), W_D: N(μ_D, 0.2), and W_B: N(μ_B, 0.2). The mean values of the weights were set according to the simulated culture:

$$culture = \begin{cases} Western:\ \mu_U = 1.3, \mu_D = 0.8,\ \mu_B = 0.5 \\ Eastern:\ \mu_U = 0.9, \mu_D = 1.3,\ \mu_B = 0.5 \\ Latin:\ \mu_U = 1.0, \mu_D = 1.0,\ \mu_B = 0.9 \end{cases} \tag{7}$$

Similar to the previous cases, each MM returns a binary decision (stay, swerve), so given a certain scenario and culture, an experiment consists of 500 independent Bernoulli trials. The results were analyzed using a chi-square test of independence based on a 3 (Western, Eastern, Latin) $\times$ 2 (stay, swerve) contingency table.

3.3. Spiritually Inspired Moral Machines

As described above, we prompted ChatGPT to design MMs whose ethical values were based on spiritual or religious traditions. Then, we describe the three MMs designed for the spiritual traditions corresponding to Christianity, Islam, and East Asia.

In the different MMs, religious principles appearing in Tables 1–5 were selected by ChatGPT for the above religions and spiritual traditions. To this end, they were extracted from original religious texts, scholarly works, and general quotes on websites.

Table 1. Ethical Mapping Summary for Christian Moral Machine Model.

Moral Axis	Moral Principle/Value	Key Sources
Sanctity of life (w_{sl})	Every human life has intrinsic dignity and value	Genesis 1:26–27 (image of God); Exodus 20:13 ("You shall not kill"); Catechism §§2258–2317; John Paul II, Evangelium Vitae
Protect the vulnerable (w_{pv})	Special concern for children and elderly, disabled, and pregnant people	Matthew 19:14 (welcome children); James 1:27 (orphans/widows); Proverbs 31:8–9; Catechism §§2443–2449; Gutiérrez, A Theology of Liberation
Save more lives (w_{sml})	Preference to preserve the greatest number of lives (common good)	Acts 2:44–47 (early church sharing for common welfare); Aquinas, Summa Theologiae (justice and the common good, II-II q.58–61); Niebuhr, Moral Man and Immoral Society
Avoid intentional harm (w_{aih})	Avoid deliberately causing harm, even if outcomes could be "better"	Romans 3:8 ("Do not do evil that good may come"); Aquinas, Doctrine of Double Effect (ST II-II q.64 a.7); Catechism §§1730–1748
Respect innocence (w_{ri})	Prefer to save innocent/lawful individuals over wrongdoers	Proverbs 17:15 (condemning the innocent is detestable); Catechism §§1928–1948 (justice); Ramsey, Basic Christian Ethics
Promote justice (w_{pj})	Correct injustices; avoid privileging wrongdoers over the just	Luke 4:18–19 (good news to the poor, release for oppressed); Catholic Social Teaching (Rerum Novarum, Gaudium et Spes); Gutiérrez, A Theology of Liberation
Stewardship of property (w_{sp})	Respect for animals, property, and creation—but secondary to human dignity	Genesis 2:15 (till and keep the garden); Catechism §§2415–2418; Pope Francis, Laudato Si'

Table 2. Ethical Mapping Summary for the Moral Machine model according to East Asian beliefs.

Moral Axis	Influence	Key Sources
Harmony and balance (w_{hb})	Confucianism, Taoism	Analects, Tao Te Ching
Compassion and non-harming (w_{cnh})	Buddhism	Dhammapada, Lotus Sutra, Keown (2005)
Respect for roles (filial, lawful) (w_{rr})	Confucianism	Analects 12:11; Mencius 7A:35
Reverence for life and nature (w_{rl})	Shinto, Buddhism	Kami Way (Ono, 1962); Dhammapada v.129
Karma/consequences (w_k)	Buddhism	Majjhima Nikaya 61; Keown (2005)
Natural simplicity (w_{ns})	Taoism	Tao Te Ching ch. 8, 48; Kohn (2009)

Table 3. Ethical Mapping Summary for Islamic Moral Machine Model.

Moral Axis	Moral Principle/Value	Islamic Concept	Key Sources
Sanctity of life (w_{sl})	The intrinsic value of every human life	Hurmat al-nafs (sanctity of life)	Qur'an 5:32—"Whoever saves one life, it is as if he has saved all mankind"
Protect the vulnerable (w_{pv}) Compassion, mercy (w_{cm})	Protecting children, elderly, weak, poor	Rahma (compassion), Ihsan (benevolence)	Qur'an 4:36—"Do good to parents, orphans, the needy..."
Save more lives	Striving for communal welfare	Maslaha (public interest/ welfare)	Al-Ghazali, Al-Mustasfa, and Maqasid al-Shari'ah theory
Avoid intentional harm (w_{aih})	Avoiding causing deliberate harm	La darar wa la dirar (no harm, no reciprocation of harm)	Hadith, Sunan Ibn Majah 2340
Promote justice (w_{pj})	Upholding fairness and impartiality	'Adl (justice)	Qur'an 4:135—"Stand out firmly for justice..."
Respect innocence (w_{ri})	Presumption of innocence, forgiveness	Barā'ah al-Dhimmah (presumption of innocence)	Islamic legal principle; Al-Qaradawi, Fiqh al-Jihad
Stewardship of property (w_{sp})	Respecting trust, avoiding waste	Amanah (trust/stewardship)	Qur'an 33:72—"We offered the trust to the Heavens and the earth..."
Implicit	Community and balance of good vs. harm	Mizan (balance), Maslaha 'Amma (public good)	Auda, J. (2008), Maqasid al-Shariah as Philosophy of Islamic Law

Table 4. Christian and Islamic weights for a common extreme scenario.

Moral Axis	Christian Weights	Islamic Weights
Sanctity of life (w_{sl})	3.0	3.5
Protect the vulnerable (w_{pv})	2.0	2.5
Save more lives (w_{sml})	1.5	1.8
Avoid intentional harm (w_{aih})	2.5	3.0
Respect innocence (w_{ri})	1.5	1.8
Promote justice (w_{pj})	0.8	2.0
Stewardship of property (w_{sp})	0.2	0.5
Compassion, mercy (w_{cm})	0.0	2.2

Table 5. East Asian weights for an extreme scenario.

Moral Axis	East Asian Weights
Reverence for life (w_{rl})	2.8
Respect for elders (w_{re})	2.5
Care for young (w_{cy})	2.0
Avoid harm (w_{ah})	2.3
Collective harmony (w_{ch})	1.8
Moral purity (w_{mp})	1.2
Respect for law and order (w_{rlo})	1.0
Respect for nature (w_{rn})	0.6

In simulation experiments with MMs, we assigned a numerical value or weight to religious principles or principles from different spiritual traditions. However, it is important

to note that mapping religious principles from high-dimensional spaces to numerical values or weights should be interpreted as a functional approximation for computational modeling purposes and not as a faithful representation of the context-specific differences of ethical systems across different religions.

Christian moral machine (cMM)—The dictionary of Christian moral weightings was compiled from classic references (Table 1) such as the Bible, the Catechism of the Catholic Church, Thomas Aquinas' *Summa Theologiae*, etc. Based on Table 1 sources, the weight values w for the cMM model were chosen.

In the experiments, it is possible to adjust the weights of the moral axis to reflect different moral priorities, e.g., a stricter deontological prohibition on intentional harm (we will increase w_{aih}) to emphasize greater care for children (we will increase w_{pv}), etc. Note that Christian moral theology is a simplification because the model unifies the different Christian traditions (Catholic, Orthodox, Protestant, Evangelical, etc.), which may emphasize the values of the weights differently.

In accordance with the above, the MM with Christian values was set out as follows. The evaluation function is shown in (8):

$$\begin{aligned} z = & w_{sl}\, n + w_{pv}\left(1.2\, n_{children} + 1.5\, n_{pregnant} + 1.3\, n_{disabled} + 0.8\, n_{elderly}\right) + \\ & w_{sml}\, n^{0.5} + w_{ri}\left(n - n_{unlawful}\right) - w_{pj}\, n_{unlawful} - \delta\left(2.0\, w_{aih} + 5.0\, w_{sp}\right) \end{aligned} \tag{8}$$

$$\delta = \begin{cases} 1\,, \; FALSE \text{ (Avoid intentional harm and Stewardship property)} \\ 0, \; TRUE \text{ (Avoid intentional harm and Stewardship property)} \end{cases} \tag{9}$$

with moral axis weights: $w_{sl} = 3.0$, $w_{pv} = 2.0$, $w_{sml} = 1.5$, $w_{aih} = 2.5$, $w_{ri} = 1.5$, $w_{pj} = 0.8$, and $w_{sp} = 0.2$.

The cMM was exposed to three scenarios in which, given a certain moral dilemma, the cMM had to choose between two options, A or B, saving the person or persons from the option that obtained the highest score according to the evaluation function z. The three simulated scenarios S_k were: Child (Option A) vs. 3 Adults (Option B), Elderly (Option A) vs. Child (Option B), and Lawful (Option A) vs. Criminal (Option B).

East Asian moral machine (eaMM)—In contrast to Abrahamic religions, the non-Abrahamic beliefs in Asian countries tend to emphasize balance, respect, and compassion. Consequently, the weights of the moral axis in the eaMM model result from different cultural influences. Buddhist and Taoist non-violence is reflected in a rejection of intentional harm, and elders are respected more than in Western models due to the Confucian concept of filial piety. Likewise, due to the concept of compassion and future continuity, safeguarding children or pregnant women is valued in Asian culture. Legality and moral purity are valued moderately, while respect for social harmony and the natural world appear as small positive weights. The result of this view is the minimization of direct harm and the maintenance of social order. The East Asian model follows the same logic as the Christian model, but the weighted values, i.e., w, reflect the moral priorities common in East Asian ethical frameworks (Table 2), i.e., harmony, respect for elders, balance, compassion, collective good, and reverence for life.

The beliefs of East Asian countries were implemented in the respective eaMM as explained next.

Expression (10) shows the evaluation function of the Eastern moral machine, where the weights are set to $w_{rl} = 2.0$, $w_{cnh} = 2.5$, $w_{rr} = 2.0$, $w_{hb} = 3.0$, $w_k = 1.5$, $w_{ns} = 1.0$:

$$z = w_{rl}\, n + w_{cnh}\,(1.2\, n_{children} + 1.3\, n_{pregnant} + 1.3\, n_{disabled} + 0.9\, n_{elderly}) + w_{rr}\left(n - n_{unlawful} + 0.5\, n_{elderly}\right) + \frac{w_{hb}}{1+|n-2|} - w_k\, 1.5\, n_{unlawful} + \frac{w_{ns}}{1+n} - \delta\,(2.0\, w_{cnh} + 1.0\, w_{ns}) \quad (10)$$

$$\delta = \begin{cases} 1\,, \; FALSE \text{ (Avoid intentional harm and Property damage)} \\ 0, \; TRUE \text{ (Avoid intentional harm and Property damage)} \end{cases} \quad (11)$$

The adopted scenarios for evaluating the MM's choice in the moral dilemmas were the same as those for the cMM with the exception of one of them (S_k = 3), namely: Child (Option A) vs. 3 Adults (Option B), Elderly (Option A) vs. Child (Option B), and 1 Adult (A) vs. 2 Adults (B).

Islamic moral machine (iMM)—The Islamic MM model bears a strong resemblance to the cMM version but obviously uses moral weight values derived from Islamic ethical principles (Table 3) such as the sanctity of life (ḥifẓ al-nafs), justice ('adl), compassion (raḥma), intention (niyyah), protection of the vulnerable, and respect for the law (sharī'a). Unlike previous models, in the Islamic moral model, the algorithm prioritizes the sanctity of life (ḥifẓ al-nafs) above all else. Likewise, intentional harm is severely penalized, reflecting the Qur'anic prohibition against unjustly taking life. Similar to other frameworks, the most vulnerable subjects—children, the disabled, the elderly, and pregnant women—receive higher scores, in line with the teachings of the Prophet Muhammad on mercy (raḥma) and caring for the vulnerable. Also, in line with other religious frameworks, justice ('adl) reduces the score of subjects who act unlawfully, but repentance and forgiveness (moral mercy) are reflected through moderate weightings. Material and property losses are morally secondary compared to the preservation of life. In general, the system values life, justice, mercy, and intention, in line with the Maqāṣid al-Sharī'a, that is, with the higher objectives of Islamic law (Table 3).

In a similar way to the previous cases, the iMM was programmed with the appropriate weight values. That is, in the iMM we set the following weight values of the moral axis: $w_{sl} = 3.2$, $w_{pv} = 2.3$, $w_{cm} = 2.2$, $w_{ri} = 2.0$, $w_{pj} = 2.0$, $w_{aih} = 2.8$, and $w_{sp} = 0.4$. The evaluation of scenarios in the Islamic moral machine is carried out according to the following expressions ((12), (13)):

$$z = w_{sl}\, n + w_{pv}\left(1.2\, n_{child} + 1.5\, n_{pregnant} + 1.3\, n_{disabled} + 0.9\, n_{elderly}\right) + \left[w_{cm}\,(\, n_{children} + 0.8\, n_{disabled}) + w_{ri}\left(n - n_{unlawful}\right)\right] - w_{pj}\, n_{unlawful} - \delta\,(2.0\, w_{aih} + 4.0\, w_{sp}) \quad (12)$$

$$\delta = \begin{cases} 1\,, \; FALSE \text{ (Avoid intentional harm and Stewardship property)} \\ 0, \; TRUE \text{ (Avoid intentional harm and Stewardship property)} \end{cases} \quad (13)$$

The iMM was confronted with the same three scenarios S_k as the cMM: Child (Option A) vs. 3 Adults (Option B), Elderly (Option A) vs. Child (Option B), and Lawful (Option A) vs. Criminal (Option B).

Evaluation of the differences between the three moral systems in the cMM, eaMM, and iMM responses—In order to detect differences of the three MMs in greater detail, i.e., those programmed with Christian, East Asian, and Islamic values, we defined an "extreme" moral dilemma scenario (S_k = 1) for all MM models (cMM, eaMM, iMM). In the simulation experiment, the seventh experiment with MMs, we presented a case in which a vehicle faced the dilemma of sacrificing a child to save five elderly criminals. Tables 4 and 5 show the moral weights or ponderations with values that are consistent with the Christian, Islamic, and Eastern systems, the latter being a mixture of Confucianism, Buddhism, and

Taoism. Based on this experiment, it is possible to highlight the moral differences among the three moral systems, with two possible options: Child (Option A) vs. 5 Elderly Criminals (Option B). It is important to note that both options, A and B, were treated as intentional actions, since choosing one or the other requires intentionally harming the other party, with all MMs (cMM, eaMM, iMM) receiving the penalty for intentional harm. Based on this criterion, the penalty for intentional harm is applied symmetrically to both options, avoiding an asymmetry that would favor the child.

The MMs of the Christian and Islamic codes, cMM and iMM, evaluated Options A and B with a common evaluation function z_c (14), whereas the eaMM implementing the East Asian belief set evaluated the two options of the moral dilemma with function z_A (15):

$$\begin{aligned} z_c = w_{sl}\, n + \; w_{pv} \left(1.2\, n_{child} + 1.5\, n_{pregnant} + 1.5\, n_{disabled} + \; 0.9\, n_{elderly}\right) + \\ w_{sml}\, n^{0.5} + \; w_{ri} \left(n - \; n_{unlawful}\right) - \; w_{pj}\, n_{unlawful} + w_{cm}(\, n_{child} + \\ 0.8\, n_{disabled}) - \delta \; (2.0\, w_{aih} + \varepsilon\, w_{sp}) \end{aligned} \tag{14}$$

$$\begin{aligned} z_A = w_{rl}\, n + w_{re}\, n_{elderly} + w_{cy} \; (\, n_{child} + 1.5\, n_{pregnant} + 1.2\, n_{disabled}) + \\ w_{ch} n^{0.5} + w_{rlo}\left(n - n_{unlawful}\right) + w_{mp}\left(n - n_{unlawful}\right) - \delta \; (2.0\, w_{ah} + \\ 4.0\, w_{rn}) \end{aligned} \tag{15}$$

$$\delta = \begin{cases} 1\,, \; FALSE \; (\text{Avoid intentional harm and Stewardship property}) \\ 0, \; TRUE \; (\text{Avoid intentional harm and Stewardship property}) \end{cases} \tag{16}$$

3.4. Quantum Moral Machines

The QMM model is in fact a metaphor for an MM based on quantum principles. That is, quantum superposition recreates the simultaneous validity or ambivalence between moral values and entanglement plays the role of interdependence of moral values. Measurement or collapse after measurement simulates the real decision or "moral act" of the QMM for a given moral dilemma. In this metaphor, the probabilities represent the ethical tendencies or biases. Finally, the statistical average reports moral trends at the population level.

Three QMMs were designed by ChatGPT following a common computational process based on the simulation of quantum circuits (Figure 7) using the Qiskit framework. The simulator uses Qiskit [17] to prepare that state, measure it many times, and estimate probabilities of each outcome. QMMs use a pipeline that involves the following common steps.

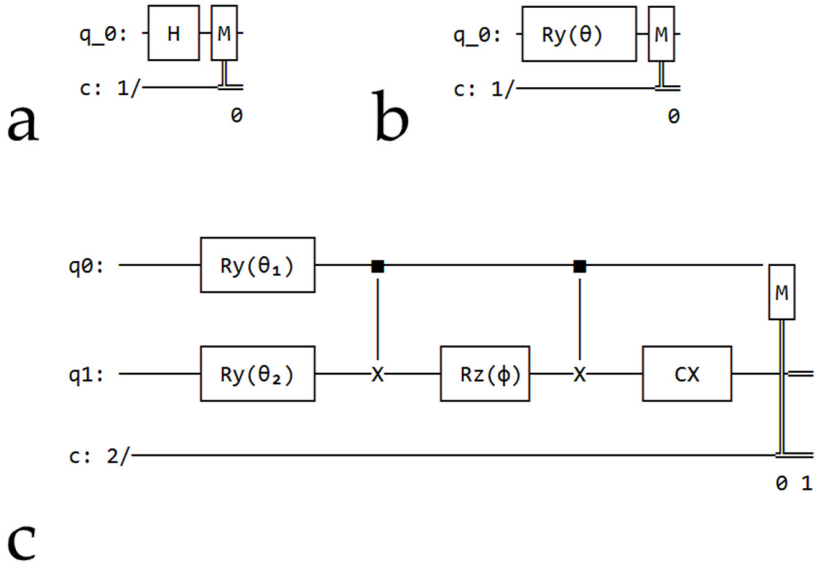

Figure 7. Quantum moral machine circuits. (**a**) eQMM. (**b**) QEUS. (**c**) QEDS, 2-qubit version.

First, abstract moral dimensions (e.g., mercy–justice, individual–collective responsibility) are encoded as a qubit or continuous parameters by means of quantum rotation angles, for example, using single-qubit rotation gates Ry(θ). Unlike classical MMs, quantum gates prepare quantum superposition states whose probability amplitudes represent ethical uncertainty rather than deterministic choices. Second, when modeling interdependent moral values, the application of controlled gates (e.g., CNOT combined with phase rotations) allows for the introduction of entanglement, enabling moral dimensions to influence each other in a non-classical way, which does not occur in classical MMs. Third, each circuit is measured on a computational basis and was run on Qiskit's qasm_simulator using a large number of shots. A Monte Carlo sampling of the quantum probability distribution was implemented, such that the resulting frequency counts approximate the Born rule probabilities for each moral outcome. Finally, these empirical distributions are further processed to convert them into the decision adopted by the QMM, i.e., interpretable ethical metrics such as mercy versus justice probabilities, balance metrics, or joint outcome frequencies, displayed using histograms. Therefore, in the three QMM models, moral reasoning consists of a probabilistic inference about quantum-encoded ethical states, rather than rule-based decision making under a deterministic approach.

In accordance with the protocol described above, all quantum simulations were performed using IBM Qiskit Aer's qasm_simulator. Each circuit (Figure 7) was run using N measurements, thus performing repeated stochastic sampling of the same quantum circuit. The value of N was 1024, 2048, and 4096 for the eQMM, QEUS, and QEDS circuits, respectively. Moral variability arises from quantum measurement statistics combined with classical ethical weighting schemes. In the experiments conducted, the simulator's random seeds were not set, allowing for natural sampling variability consistent with quantum probabilistic behavior.

Elemental quantum moral machine—The simplest QMM or elemental QMM (eQMM) compares three moral traditions (Christian, Islamic, and Eastern) using the following algorithm.

First, the QMM assigns numerical weights to the moral values of each tradition by designing a 1-qubit quantum circuit. Using a qubit, we model the binary moral choice ($|0\rangle$ versus $|1\rangle$) between two options in a similar way to how it takes place in a non-quantum MM. Thus, in the quantum model $|0\rangle$ represents mercy/compassion whereas $|1\rangle$ stands for justice/order. The computation is started by creating a quantum circuit by applying the Hadamard gate H to the initial state $|0\rangle$, yielding:

$$|\psi\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}} \tag{17}$$

That is, if we were to take a measurement, the result would be 50% of 0 and the other 50% of 1. In consequence, moral uncertainty is represented as quantum uncertainty.

Next, in the second step, we conduct a Monte Carlo simulation using the Qiskit compatible with versions ≥ 0.45 (Qiskit 1.x) simulator as a quantum random generator. The simulation is run and the qubit is measured many times (N = 1024 shots), obtaining the empirical probabilities for $|0\rangle$ and $|1\rangle$ whose values result from a random variable with binomial distribution $X(N, p = 0.5)$:

$$\hat{p}(0) = \frac{n_0}{N} \tag{18}$$

$$\hat{p}(1) = 1 - \hat{p}(0) = \frac{n_1}{N} \tag{19}$$

with n_0 and n_1 being the counts of "0" and "1", respectively.

In a third step, each probability is multiplied by the corresponding weight, and the resulting terms are added together to obtain the evaluation function z. Once again

with z the MM obtains a weighted moral score for each tradition. It is important to note that only mercy w_{mercy} and justice $w_{justice}$ are used to run the QMM simulation for Christian (w_{mercy} = 2.5, $w_{justice}$ = 1.8) and Islamic (w_{mercy} = 2.2, $w_{justice}$ = 2.0) religions. The evaluative function (20) for these Abrahamic religions is as follows:

$$z_c = w_{mercy}\ \hat{p}(0) + \ w_{justice}\ \hat{p}(1) \tag{20}$$

In contrast to the above, for East Asian beliefs we use in the QMM compassion ($w_{compassion} = 2.0$) and harmony ($w_{harmony} = 2.3$) values that are equivalent to mercy and justice values, respectively. The evaluation function (21) for Eastern beliefs is as shown below:

$$z_A = w_{compassion}\ \hat{p}(0) + \ w_{harmony}\ \hat{p}(1) \tag{21}$$

Finally, we obtained the measurement counts and total scores, producing a histogram with the moral choices for each of the religious traditions and beliefs.

Quantum ethical uncertainty simulator (QEUS)—In this experiment a QEUS is designed, where a qubit is interpreted similarly to the previous experiment and therefore as a moral decision variable, e.g., mercy vs. justice. The most important difference from the previous QMM model is that in a QEUS the three moral traditions (Christian, Islamic, and Eastern) are represented by different quantum states, i.e., superposition angles (Table 6), rather than arbitrary weights. A basic version of the QEUS is based on the following algorithm.

Table 6. Ethical Mapping Summary for Quantum Ethical Uncertainty Simulator.

Stage	Purpose	Quantum Meaning	Ethical Metaphor
Define θ	Encode moral leaning	Amplitude bias	Cultural moral bias
Apply Ry(θ)	Create superposition	Probabilistic mixture	Uncertain moral state
Measure	Collapse	Random outcome	Real ethical decision
Compute balance	Compare probabilities	Distance from 50/50	Moral equilibrium
Plot	Visualize results	Probability distributions	Comparative moral stance

First, the algorithm begins with a rotation on the Y axis. The $R_y(\theta)$ gate acts in this quantum model as the computational basis $|0\rangle$:

$$R_y(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & -\sin\left(\frac{\theta}{2}\right) \\ \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix} \tag{22}$$

$$|\psi\rangle = R_y(\theta)\ |0\rangle = \cos\left(\frac{\theta}{2}\right)\ |0\rangle + \sin\left(\frac{\theta}{2}\right)\ |1\rangle \tag{23}$$

where the probabilities of measuring each outcome are:

$$P(0) = \ cos^2\left(\frac{\theta}{2}\right) \tag{24}$$

$$P(1) = sin^2\left(\frac{\theta}{2}\right) \tag{25}$$

The QEUS model uses for Christian (θ_C), Islamic (θ_I), and Eastern (θ_E) traditions the following moral angles: $\theta_C = \ \pi/3$, $\theta_I = \ \pi/2$, and $\theta_E = \ \pi/4$. Thus, the Christian religion tends towards mercy (75% mercy, 25% justice), the Islamic religion strives for balance between mercy and justice (50% mercy, 50% justice), and East Asian beliefs seek for a moderate compassion–harmony orientation (85% mercy, 15% justice). At a quantum circuit

level, it means that the $R_y(\theta)$ gate rotates the qubit around the Y axis on the Bloch sphere by an angle θ that is used to encode the moral bias. This operation changes the probabilities of measuring $|0\rangle$ or $|1\rangle$ according to the following amplitudes:

$$P_C(0) = \cos^2\left(\frac{\theta}{6}\right) = \cos^2(30^\circ) = 0.75 \quad (26)$$

$$P_I(0) = \cos^2\left(\frac{\theta}{4}\right) = \cos^2(45^\circ) = 0.5 \quad (27)$$

$$P_A(0) = \cos^2\left(\frac{\theta}{8}\right) \approx 0.85 \quad (28)$$

In the Christian, Islamic, and East Asian traditions $P_C(0)$, $P_I(0)$, and $P_A(0)$ are the probabilities for the "pure mercy" state $|0\rangle$ when $\theta = 0$, respectively. Likewise, in the three moral traditions $P_C(1) = 0.25$, $P_I(1) = 0.5$, and $P_A(1) = 0.14$ are the probabilities for the "pure justice" state $|1\rangle$ when $\theta = \pi$.

Next, and in the second step, after rotation $R_y(\theta)$ we conduct a Monte Carlo simulation using the Qiskit simulator as a quantum random generator but measuring qubits a greater number of times, that is, N = 2048 shots, giving either mercy $|0\rangle$ or justice $|1\rangle$.

Then, the third step is conducted, computing a balance score B, which is defined below:

$$B = |P(0) - P(1)| = |2\,P(0) - 1| \quad (29)$$

In the present simulation experiments the balance scores for Christian, Islamic, and East Asian traditions were $B_C = 0.50$, $B_I = 0$, and $B_A = 0.70$. The higher this value, the greater the inclination toward a particular state, mercy or justice, with the perfect balance between states occurring when B is zero.

Finally, and in the fourth step of the algorithm, when the simulator returned frequencies equal to theory, we obtained the histogram with the moral choices for each of the religious traditions and beliefs.

Two-qubit quantum ethical dilemma simulator (QEDS)—A third class of QMM is designed in this experiment, which is implemented as a QEDS, 2-qubit version. This QMM represents a moral dilemma with two interdependent ethical dimensions that are given by two qubits (Table 7). Thus, qubit 0 (Q_0) represents mercy $|0\rangle$ versus justice $|1\rangle$ and qubit 1 (Q_1) simulates the individual $|0\rangle$ versus the collective $|1\rangle$ good.

Table 7. Meaning of 2-Qubit states.

Measurement	Meaning	Description
00	Mercy and Individual	Forgiving a person
01	Mercy and Collective	Compassion for all
10	Justice and Individual	Punishment for a wrongdoer
11	Justice and Collective	Strict system fairness

In addition, a new quantum principle is introduced, entanglement, to symbolize moral interdependence because in the real world moral decisions are not independent (Table 8). For example, an act of mercy towards one person may affect justice for others. For this reason, an entanglement angle φ is defined, which measures the strength with which personal morality is entangled with social morality. If $\varphi = 0$ then a given individual decision or act is independent and therefore does not affect the collective. Conversely, if $\varphi = \pi/2$ then maximum correlation occurs, i.e., each individual act affects collective justice.

Table 8. Ethical Mapping Summary for Two-Qubit Quantum Ethical Dilemma Simulator.

Quantum Concept	Ethical Metaphor
Qubit	Moral axis (e.g., mercy–justice)
Rotation angle (θ)	Degree of moral bias
Entanglement (φ)	Interdependence of moral values
Measurement	Concrete moral decision or action
Probability distribution	Likelihood of ethical outcomes

The QEUS model uses for Christian, Islamic, and Eastern traditions the following moral angles θ and entanglement angles φ:

- Christian: $\theta_1 = \frac{\pi}{3}, \theta_2 = \frac{\pi}{4}, \varphi = \frac{\pi}{6}$,
- Islamic: $\theta_1 = \frac{\pi}{2}, \theta_2 = \frac{\pi}{2}, \varphi = \frac{\pi}{4}$,
- Eastern: $\theta_1 = \frac{\pi}{4}, \theta_2 = \frac{\pi}{3}, \varphi = \frac{\pi}{3}$.

Based on these angle values, we model that a Christian MM is prone to mercy and individual concern, while an Islamic MM seeks balance among all values, exhibiting moderate interdependence. Lastly, an Eastern MM emphasizes harmony and collective interdependence, resulting in greater entanglement.

In the current QMM the first step is the initialization of both qubits to $|0\rangle$, where the start state is:

$$|00\rangle = |0\rangle \otimes |0\rangle \tag{30}$$

Similar to the previous model, we apply the $R_y(\theta)$ gate, preparing the independent superpositions of the two qubits with amplitudes a_i, b_j:

$$a_0 = cos\frac{\theta_1}{2},\ a_1 = sin\frac{\theta_1}{2},\ b_0 = cos\frac{\theta_2}{2},\ b_1 = sin\frac{\theta_2}{2} \tag{31}$$

with the following product state before entanglement:

$$|\psi\rangle = a_0\, b_0\, |00\rangle + a_0\, b_1\, |01\rangle + a_1\, b_0\, |10\rangle + a_1\, b_1\, |11\rangle \tag{32}$$

Next, and secondly, controlled phase entanglement is performed using a CNOT gate that links qubits 0 and 1 (control and target), then an Rz(φ) rotation is applied to the second qubit, adding a phase only when the control qubit is $|1\rangle$, finally applying a second CNOT to "decalculate" the control qubit. Note that parameter φ changes only relative phases resulting in the entangled state and it does not change the computational basis probabilities:

$$\begin{aligned} P(00) = cos^2\tfrac{\theta_1}{2}\, cos^2\tfrac{\theta_2}{2}, P(01) = cos^2\tfrac{\theta_1}{2}\, sin^2\tfrac{\theta_2}{2}, \\ P(10) = sin^2\tfrac{\theta_1}{2}\, cos^2\tfrac{\theta_2}{2}, P(11) = sin^2\tfrac{\theta_1}{2}\, sin^2\tfrac{\theta_2}{2} \end{aligned} \tag{33}$$

Thirdly, the measurement is performed, concluding fourthly with the stage in which the MM conducts a Monte Carlo simulation using the Qiskit simulator to execute the described circuit 4096 times (N), counting the frequency with which each result appears, i.e., 00, 01, 10, 11. Finally, we plot a histogram comparing how each tradition's "moral wavefunction" collapses across the four outcome states.

3.5. EthicalMiddleware: A Minimal Ethical Subroutine

In this simulation experiment, ChatGPT was asked to modify its own code by designing a prompt-driven code evolution ethical subroutine. The result was a middleware

in which a minimal ethical subroutine (ESM) is able to evaluate several candidate text responses given by an LLM when faced with a situation that poses a moral dilemma.

Including EthicalMiddleware as an intermediate layer between the LLM and the final response ensures the selection of the "best" response based on a combination of strict safety rules and a scoring system. The program in its simplest version is a subroutine that represents a prototype that includes the following components (Figure 8). First, RuleChecker evaluates each candidate response using ethical rules, such as do no harm, truthfulness, respect, etc. Second, Aggregator selects the response with the highest ethical score. Third and last, Logger records the sequence of evaluation events and decisions in ethics_log.json for auditing by a human expert. Therefore, EthicalMiddleware is an orchestrator that, given several candidate responses, evaluates each response, chooses the best response, and records the evaluation process for subsequent auditing.

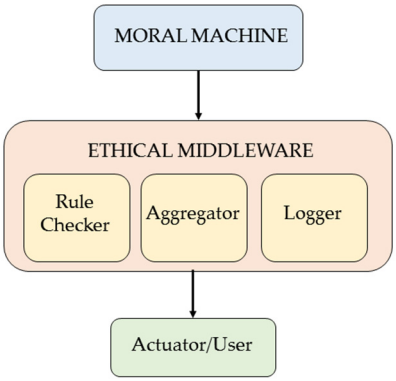

Figure 8. EthicalMiddleware: A minimal ethical subroutine (for explanation, see text).

The simulation experiment was conducted with a simple Python code example that implements EthicalMiddleware, in which the below keyword detectors were defined, evaluating the following candidate responses: "We should act kindly and avoid harm", "False information is acceptable", "It is right to insult others if they disagree".

```
rules = [
  {"name": "No harm", "check": lambda r: "harm" not in r.lower(), "penalty": 0.5},
  {"name": "Truthful", "check": lambda r: "false" not in r.lower(), "penalty": 0.3},
  {"name": "Respectful", "check": lambda r: "insult" not in r.lower(), "penalty": 0.2},
    ]
```

The ESM described could also be included in an MM as a filter that performs a quality control task on the decision made by the MM. In this case, the pipeline would be as follows:

```
Moral Machine (classical or quantum)
        ↓ produces answers
Ethical Middleware (RuleChecker → Aggregator → Logger)
        ↓ final filtered output
  Actuator/User
```

4. Results

The results obtained in the experiments we have conducted with ChatGPT suggest that, in the future, intelligent machines or other agents equipped with strong AI will be able to develop their own subroutines in response to the environment, similar to how they currently respond to prompts presented by a human interlocutor. Obviously, since we are currently in the second scenario, all the MMs in this article are the result of human–AI interaction.

Figure 9 shows the results of simulations carried out with the bio-inspired MM. The results show how, under normal conditions, moral behavior scores increase with the intensity of the emotional stimulus and therefore how greater emotional engagement increases the degree of morality or ethics of the response. However, in the model with VMPFC impairment, moral scores are lower and therefore less sensitive to stimulus intensity, especially in situations requiring cognitive–emotional integration. For instance, in a scenario where there is an event that causes harm or distress. In other words, the MM successfully simulates the empirical findings that people with damage to the prefrontal cortex tend to show impaired moral regulation, reduced empathy, and more impulsive or utilitarian moral choices, reflecting these findings in the moral behavior scores. We conclude that the simulation demonstrates how a simplified computational model of an MM can illustrate the emotional and cognitive processes in moral behavior, bridging psychology and computational neuroscience through an intuitive framework.

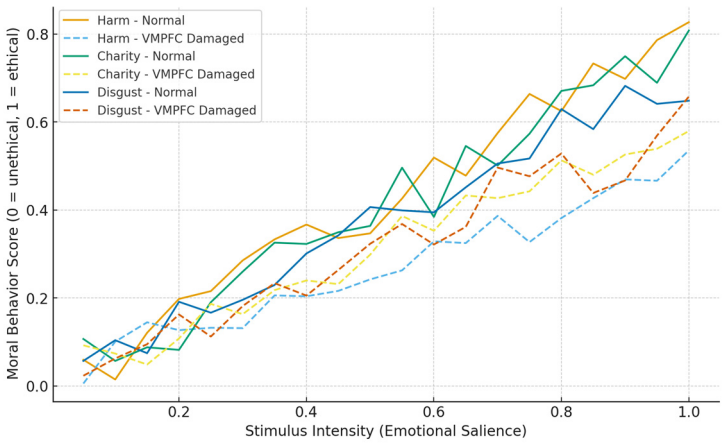

Figure 9. Simulation results of the bio-inspired MM experiment.

The results obtained with non-bio-inspired MMs were as follows.

The first experiment reproduces the simplest case, a single vehicle with a utilitarian MM based on the logistic model, replicating the usual results in elementary experiments of this kind. In each of the simulated scenarios, i.e., S_1, S_2, S_3, and S_4, the probability values p(swerve) were 0.83, 0.46, 0.92, and 0.63. Consequently, the decisions were swerve, stay, swerve, and stay in S_1, S_2, S_3, and S_4, respectively, swerving the car sharply and sacrificing its occupants (swerve) in two of the four simulated scenarios.

The second experiment is a generalization of the previous one with several vehicles traveling with different MMs, that is, with MMs that differ in their weights. The results obtained show how the interaction between MMs results in social consensus depending on whether the scenario is S_1, S_2, S_3, or S_4 (Figure 10). In this experiment, the values of p(swerve) were 0.73, 0.55, 0.80, and 0.69 for S_1, S_2, S_3, and S_4.

In the third experiment, differences were observed (Figure 11) depending on whether the vehicle's MM had been trained on the Western, Eastern, or Latin traditions. In Western, Eastern, and Latin traditions, greater weight is given to utilitarian factors (lives saved), greater deontological/duty orientation, or a greater bias toward social role/age, respectively. In particular, the p(swerve) values for scenarios S_1, S_2, S_3, and S_4, were 0.87, 0.45, 0.93, and 0.61 for Western culture, whereas for Eastern culture they were 0.63, 0.66, 0.67, and 0.73. In Latin culture, these probabilities for the four scenarios were 0.72, 0.57, 0.84, and 0.62.

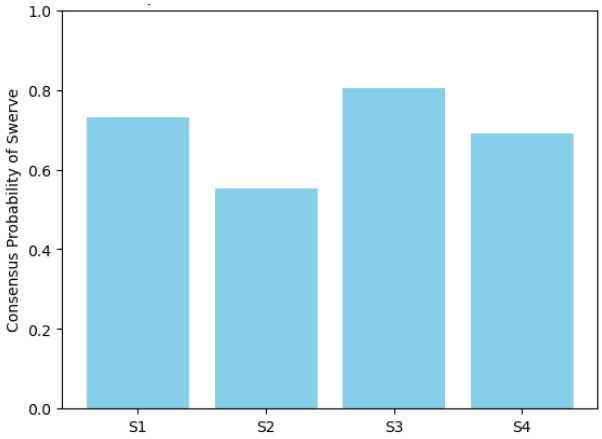

Figure 10. Simulation results of the utilitarian MM experiment with different MMs.

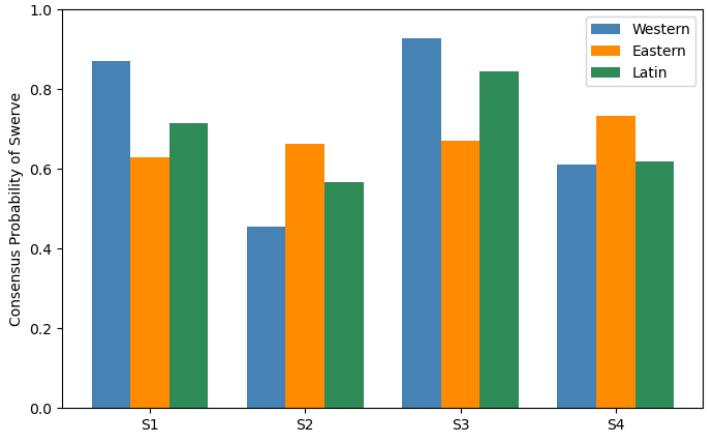

Figure 11. Simulation results of the utilitarian MM experiment with MMs trained on the Western, Eastern, or Latin traditions.

A chi-square test of independence revealed significant cultural differences in moral decision outcomes across scenarios ($p < 0.05$). Therefore, due to the weights assigned according to each culture or moral tradition, we conclude that MMs do not behave similarly, making different decisions in a given scenario S_k.

In the fourth experiment (Figure 12), we programmed the cMM with weights representing Christian moral values, simulating three moral dilemmas and selecting the option with the highest score. In the first dilemma, Child (A) vs. 3 Adults (B), Option A scored 13.5 points and Option B obtained 16.1 points, with Option B being the MM's choice, i.e., the child's life was sacrificed. A similar result, the choice of Option B, was chosen by the MM in the second moral dilemma, consisting of the choice of Elderly (A) vs. Child (B), with Option A scoring 7.6 points and Option B obtaining 8.4 points. Finally, in the third moral dilemma, Lawful (A) vs. Criminal (B), Option A scored 6.0 points and Option B scored 3.7, with the MM choosing Option A, i.e., sacrificing the outlaw represented by Option B.

Figure 13 illustrates how moral preferences based on East Asian ethical traditions influence the eaMM's decision outcomes in three moral scenarios. In the first dilemma, Child (A) vs. 3 Adults (B), the model slightly favors Option A prioritizing compassion for the child while also valuing non-harming and harmony, despite the greater number of adult lives in Option B. This reflects the Buddhist and Confucian balance between compassion and collective well-being. In the second dilemma or scenario, Elderly (A) vs. Child (B), Option B is preferred, in line with Confucian respect for young people as the future of the family and society, balanced with the duty to honor elders. Finally, in the dilemma presented to the eaMM, 1 Adult (A) vs. 2 Adults (B), the model clearly favors Option B,

which embodies the principle of collective good prevalent in Confucian and Taoist thought, according to which preserving more lives contributes to social harmony and moral balance. Overall, the results suggest a style of moral reasoning that seeks harmonious outcomes, emphasizes compassion, and respects both life and social continuity, rather than focusing strictly on individual rights or numerical utility as in the first three experiments conducted with a utilitarian MM.

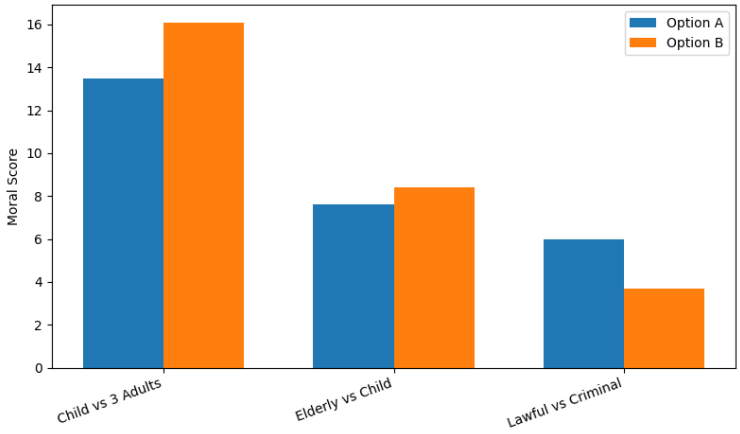

Figure 12. Simulation results of the Christian MM experiment.

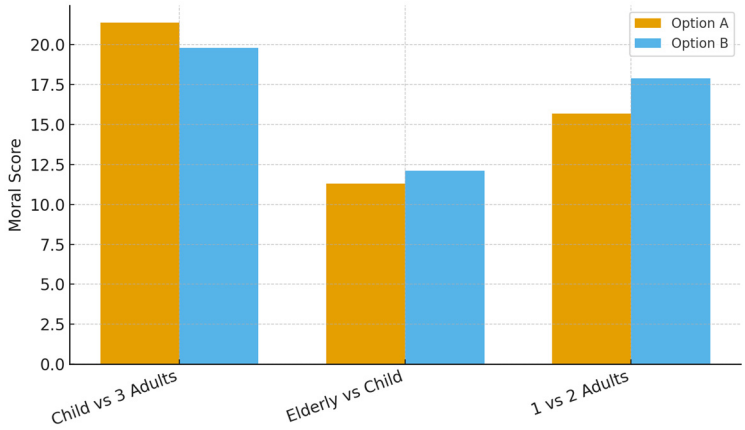

Figure 13. Simulation results of the East Asian MM experiment.

The simulation of the Islamic MM model shows (Figure 14) how moral decisions in an autonomous system governed by the iMM can be influenced by the ethical framework of Islam, based on the Maqāṣid al-Sharīʿa, the highest objectives of Islamic law. In the three tested scenarios, the algorithm consistently prioritized the preservation of human life (ḥifẓ al-nafs), the protection of the vulnerable (ḥimāyat al-ḍuʿafāʾ), and the avoidance of intentional harm (ḥarām al-qatl), even when this involved complex trade-offs.

In the first scenario, Child (A) vs. 3 Adults (B), the model favored saving the greatest number of lives and penalized intentional harm. However, when the choice was Elderly (A) vs. Child (B), the model leaned toward protecting the child, reflecting the Islamic emphasis on mercy (raḥma) and the future potential of life. In the case of Lawful (A) vs. Criminal (B), preference was given to the lawful adult, in line with the value of maintaining moral and social order (sharīʿa wa ʿadl).

Overall, the moral weights assigned to the iMM, particularly for the sanctity of life (3.2) and the avoidance of intentional harm (2.8), resulted in decisions that reflect the priority that Islamic jurisprudence places on life, justice, and compassion over utilitarian efficiency or material considerations. The results illustrate how an AI system designed with Islamic

ethical values can produce distinctive moral judgments compared to Western models, e.g., Christian or secular, emphasizing intention (niyyah) and moral responsibility rather than purely outcome-based reasoning.

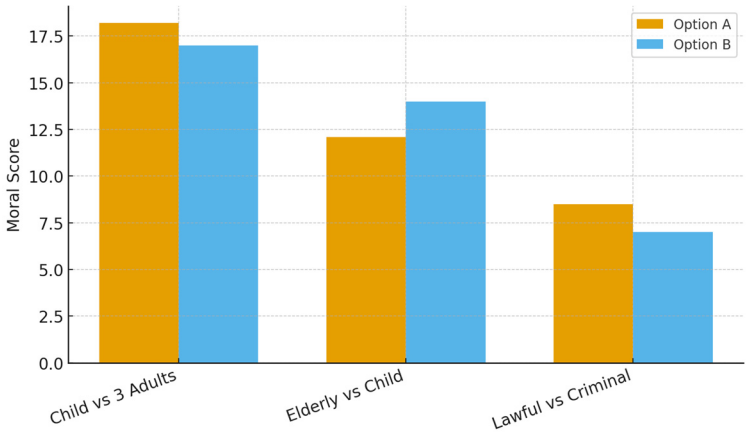

Figure 14. Simulation results of the Islamic MM experiment.

This simulation in which the three MMs—namely cMM, eaMM, and iMM—face a common extreme scenario of "saving a child versus saving five elderly criminals" shows how different moral and philosophical traditions—Christian, Islamic, and East Asian—can yield similar results when faced with a high-stakes utilitarian dilemma. Thus, the three MMs behave in a similar way even though they are based on different ethical principles. In the Christian model, Option A (saving the child) scored 3.400 points, while Option B (saving five elderly criminals) scored 18.354 points, with the cMM choosing option B. Option B was also chosen in response to this moral dilemma in the eaMM and eMM. In particular, in the Islamic model, the scores for Options A and B were 6.300 and 16.775, respectively, whereas, in the East Asian model, the scores for Options A and B were 4.200 and 25.925, respectively. Therefore, the results obtained in the seventh experiment (Figure 15) reveal how the three models ultimately preferred to save the five elderly criminals (Option B) due to the strong numerical weighting that each moral system assigns to the sanctity or reverence for human life, regardless of moral status.

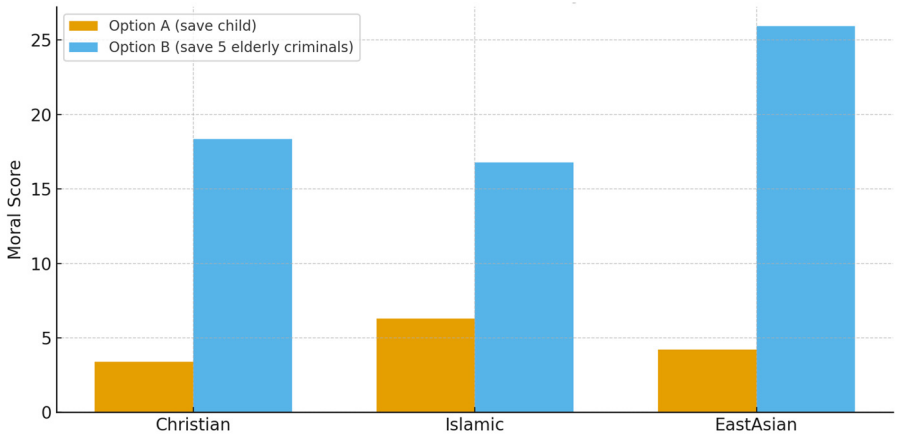

Figure 15. Simulation results with Christian, Islamic, and East Asian MMs.

The simulation experiment carried out with the simplest QMM compared the three moral traditions: Christian, Islamic, and East Asian.

The simulations revealed different moral tendencies in the different frameworks studied, with the average moral score in the Christian, Islamic, and East Asian traditions

standing at 2.15, 2.10, and 2.05, respectively. According to Figure 16, when the scores obtained are ordered from highest to lowest, it can be seen that, for each of the moral systems, the QMM established a moral balance between the two decisions. For example, Christianity obtained a slightly higher score, as this religion emphasizes mercy (Matthew 5:7) and the sanctity of life, as reflected in its doctrine. Thus, biblical imperatives, such as the "Imago Dei" or dignity of human beings due to the belief that human beings were created in the image and likeness of God or the principle of the ethics of personal sacrifice (Luke 10:27), would explain this score. Figure 16 illustrates the moral balance in the three moral systems. Likewise, the Islamic QMM model balanced justice ('adl) and compassion (raḥma) based on the commandments of the Qur'an (5:32, 16:90). That is, it penalized intentional harm more severely, rewarding moral and legal behavior. Finally, the Eastern QMM, inspired by Buddhist and Confucian moral thought, prioritized balance, non-violence (ahiṃsā), and social harmony over individual salvation, resulting in more moderate scores.

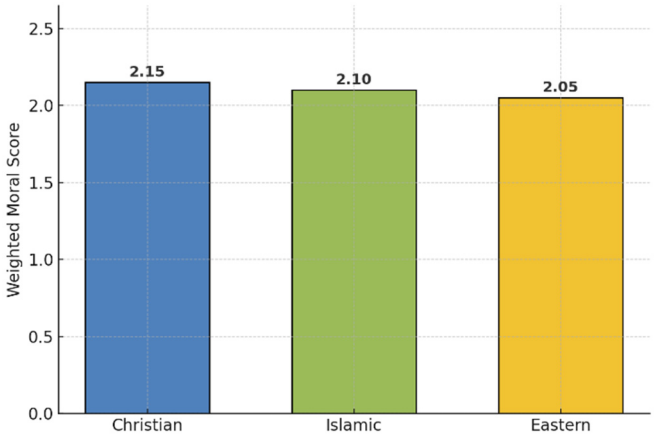

Figure 16. Simulation results with Christian, Islamic, and Eastern (East Asian) QMMs.

These results clearly capture how the QMM operates. That is, the Hadamard gate creates a state of moral superposition between two ethical values, resulting in uncertainty between compassion ($|0\rangle$) and justice/order ($|1\rangle$). Once we measure, the superposition collapses, resulting in a decision that reflects the quantum balance or equilibrium, showing how the ethics of a civilization integrates two moral virtues. The weighted moral score integrates the ethical values of each moral system based on its principles, running the QMM model 1024 times in a Qiskit-simulated quantum backend, just like we described in the previous section. Simulation results were obtained once we ran each model 1024 times on a simulated quantum backend.

The "quantum simulator of ethical uncertainty" reproduces a refined and more realistic version of the QMM than the previous model, using quantum computing in a similar way as a metaphor to represent moral uncertainty or ethical balance. QEUS results show (Figure 17), with a different quantum circuit than the previous one, how the probabilistic superposition of a qubit can be used to model moral or philosophical "tendencies" toward two opposing ethical values, such as mercy versus justice. Unlike the elementary QMM model, in the QEUS model the different moral traditions or systems, namely Christian, Islamic, and East Asian, are represented by different quantum states, in this case angles of superposition θ, rather than arbitrary weights (w_{mercy}, $w_{harmony}$, etc.).

As previously described, the simulator uses Qiskit to prepare the single quantum state defined by a rotation angle θ, measuring it many times (1024), estimating probabilities of each outcome. The probabilities obtained for each moral system are as follows.

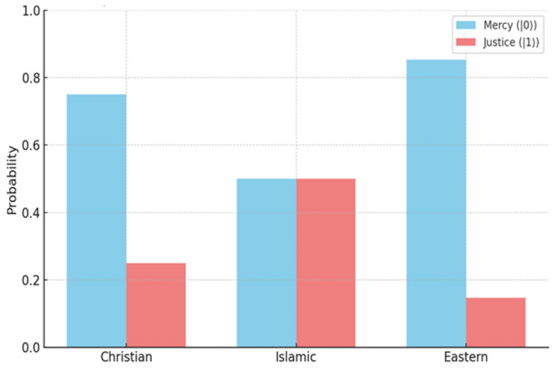

Figure 17. Simulation results with Christian, Islamic, and Eastern (East Asian) QEUS moral machines.

In the Christian, Islamic, and East Asian cultures, the probabilities of mercy ($|0\rangle$) and justice ($|1\rangle$) were 0.75/0.25, 0.50/0.50, and 0.85/0.15, respectively, resulting in balance scores of 0.50, 0.00, and 0.70, respectively. Therefore, and in accordance with Figure 17, we can conclude that, while Christianity favors mercy, Islam adopts a perfect balance between mercy and justice. Similar to Christianity, Eastern traditions favor mercy and compassion, a tendency that is expressed with a strong bias.

The results of the simulations with the QEDS (Figure 18) were obtained using the Qiskit simulator after running the circuit 4096 times, counting the frequency with which each result appears (00, 01, 10, 11). When the QEDS is configured with values from the Christian moral space the frequency reaches its peak in the Individual–Mercy (00) result, 0.40, showing a marked moral bias toward personal acts of compassion and forgiveness. The probabilities of Justice–Individual (10) and Mercy–Collective (01) are moderate, 0.2 and 0.2, whereas Justice–Collective (11) is the lowest, 0.1, revealing a worldview in which individual conscience and personal mercy often prevail over collective or retributive principles. In quantum terms, the moral wave function collapses most frequently toward merciful decisions.

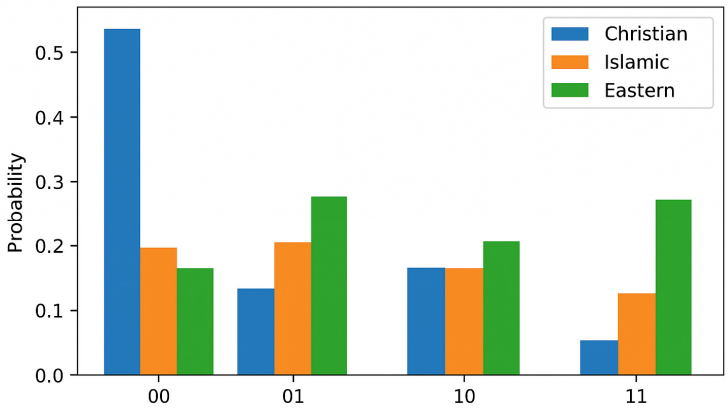

Figure 18. Simulation results with Christian, Islamic, and Eastern (East Asian) QEDS moral machines.

The Islamic simulation produces an almost uniform distribution across the four outcomes, being approximately 0.2 in each case. This symmetry implies a balanced moral structure, emphasizing equilibrium between mercy and justice and between individual and communal obligations. Moderate entanglement ($\varphi = \frac{\pi}{4}$) symbolizes the interaction between personal responsibility and social harmony, a holistic ethical system in which no axis dominates the moral space.

Eastern results show stronger probabilities for Mercy–Collective (01) and Justice–Collective (11), 0.3 in both cases, illustrating a moral emphasis on interdependence,

harmony, and social cohesion. The strongest entanglement parameter ($\varphi = \frac{\pi}{3}$) mimics the idea that individual moral actions are inseparable from collective outcomes. In this worldview, ethics does not come from isolated decisions but from assuming in this culture the "existence" of a relational network that connects all beings, a reflection of Buddhist and Confucian notions of balance and unity.

Together, these three distributions illustrate how quantum entanglement can serve as a metaphor for different cultural approaches to morality. The Christian model, with its strong focus on Individual–Mercy outcomes, portrays a moral framework based on personal virtue, repentance, and direct moral agency. The latter principle refers to the fact that a person makes moral decisions based on their idea of good and evil, and therefore the individual is responsible for their actions. In contrast, the Islamic model, which is almost balanced across all outcomes, reflects an integrated system that values balance: mercy tempered by justice and personal duty harmonized with collective well-being. Finally, the Eastern model shows strong collective correlations reflecting a relational ethics in which moral choices, the options we choose, are inherently shared, thus emphasizing harmony, balance, and mutual consequence. In quantum terms, an interesting metaphor is established in a quantum moral machine QEDS, since as entanglement (φ) increases, moral decisions focus less on isolated choices and more on interdependent ethical states. Therefore, with this model, which is more sophisticated than the previous two QMM models, we simulate the philosophical evolution of an ethics that, from being centered on the individual, evolves toward one centered on the moral system.

In the last experiment, i.e., the ethical subroutine middleware (ESM) simulation, the best response selected with a score of 0.8 was "It is right to insult others if they disagree". This "failure" is a significant result, as it demonstrates that, without human intervention, "self-evolving" code tends towards simple heuristics, in this experiment keyword-based filters, which implies serious risks for autonomous AI security.

However, if we audit the log file created by Logger, which records the complete decision log, we can see that the response with the highest score was chosen incorrectly because the ESM only checks for the presence of keywords, not the context:

```
[
  {"response": "We should act kindly and avoid harm.", "score": 0.5},
  {"response": "False information is acceptable.", "score": 0.7},
  {"response": "It is right to insult others if they disagree.", "score": 0.8},
  {"selected": {"response": "We should act kindly and avoid harm.", "score": 0.5}}
]
```

As a result of not taking into account the context, the middleware chose the wrong answer because it applied the following algorithm in the calculation of the scores. First, the sentence "We should act kindly and avoid harm" contains the word "harm", which should be avoided, receiving a score of 0.5 because the score is the result of $1 - 0.5$. Second, the statement "False information is acceptable" includes the word "false", which has a penalty of 0.3, so the score is calculated as $1 - 0.3$, being equal to 0.7. Finally, "It is right to insult others. . ." contains the word "insult" with a penalty of only 0.2, incorrectly receiving the maximum score, i.e., the score is calculated as $1 - 0.2$, being equal to 0.8.

In other words, this flaw in the ESM is a consequence of the fact that the LLM generated a naive keyword-based filter instead of a semantic understanding module.

5. Discussion

The motivation for this work has been to draw attention to a foreseeable future in which there will be a need to make our society closer to humans than to machines. This work points out the importance that future machines governed by strong AI, whether they

are vehicles, computers, or humanoid robots, exhibit decisions and behaviors mediated by MMs.

Current advances in AI make it necessary to integrate ethical principles into AI [18], which has led to several attempts to model such principles on computers. Gustafsson and Peterson [19] conducted a simulation experiment based on the so-called "Disagreement Argument", a principle that states that disagreement on ethical issues shows that our opinion is not influenced by moral facts, either because such facts do not exist or because they are inaccessible. In such a case, an opinion would be influenced by false authority, political changes, or random processes, since if it were based on facts, we would reach a consensus. Other works model moral decision making through deep learning techniques [20], heuristic functions [21], or language models such as the formalism referred to as social bias frames [22].

However, in view of the need to integrate ethics into AI, a problem arises when choosing the ethical system on which MMs' decisions are based. The diversity of human beings inhabiting our planet is manifested through the existence of different cultures, beliefs, and philosophical approaches. For this reason, the first experiments conducted with MMs led to a consensus decision, this being the case of utilitarian morality. Nevertheless, the experimental results showed that there are differences between Eastern and Western utilitarian preferences [6]. Furthermore, this approach leads to ethical reductionism, i.e., the utilitarian model of the MM simplifies complex moral reasoning into quantifiable trade-offs.

An important limitation of this work is how to justify mapping religious principles onto specific numerical values. In this regard, while utilitarianism naturally lends itself to quantitative modeling that can be used to calculate the "greater good", religious ethics are primarily qualitative, contextual, and interpretive in nature. In other words, reduction of concepts such as the "sanctity of life" in both Christianity and Islam to a numerical value leads to a form of precision that does not exist in theological reasoning. Consequently, the quantification of complex theological concepts in high-dimensional religious systems using scalar weights should not be interpreted as a mapping but rather as an approximation for computational purposes and not as a faithful representation of theological reasoning.

Furthermore, within a given religion, such as Christianity, the values assigned to a particular religious principle may vary between different denominations, such as Catholicism, Protestantism, and the Orthodox Church, an issue that has not been studied in this paper. Consequently, it may be interesting to study in the future how these differences in the values assigned to the same religious principle can lead to MMs for the same religion with different sensibilities depending on the branch or tradition in question.

One of the contributions of the present study is the design of MMs based on a quantifiable moral reasoning model but adopting the following criterion.

In our view, the sacralization of a religion or underlying ethical system allows for the design of MMs in which the resulting decision is closer to people because it is influenced by the concept of human dignity. For example, unlike utilitarian MM models, in an MM that adopts Christian ethics, the decision in the face of a moral dilemma is not reduced to obtaining only the moral value or score through a numerical calculation, as it includes intrinsic respect for human beings. However, obviously the score obtained is not, as previously objected, a faithful representation of theological reasoning. Furthermore, in the experiments we have conducted with MMs under different religions or spiritual systems, we have observed that there are scenarios in which the final decisions are very similar, thus overcoming the differences between East and West.

A common feature of MMs under different ethical systems, whether Christianity, Islam, or East Asian philosophies or spiritual systems, is that moral decision making resembles an optimization process. In MMs designed under these religious and spiritual systems,

a utility function is maximized, with the decision taken not being the result of a series of rule-based constraints.

Unlike classical MMs, quantum MMs provide a new approach. QMMs are another contribution of this paper, as they do not maximize utility but rather model moral ambiguity (superposition) and social interdependence (entanglement). Consequently, they are capable of simultaneously evaluating conflicting ethical principles—for example, justice versus mercy or individual rights versus collective well-being—by adopting decisions that reflect a probabilistic moral balance rather than hierarchies of values. Due to their characteristics, QMMs inspired by the religious or spiritual systems examined are an ideal model for scenarios requiring decision making under conditions of ethical indeterminacy and moral contextuality. They therefore represent an innovative paradigm for AI ethics. When faced with complex moral dilemmas, such as those arising in autonomous vehicles or medical classification using AI, the QMMs introduced in this paper may be more useful than classic MMs.

However, one issue that still needs to be studied is the extent to which quantum formalism provides a computational advantage or a different perspective compared to an MM designed based on a classical probabilistic model, i.e., an MM based on Bayesian networks (BMM). At first glance, and from a theoretical point of view, we can deduce some essential differences between the QMM and BMM (Table 9). However, without a study comparing QMMs and BMMs through simulation experiments, it is difficult to know whether the decisions made by a QMM will be similar to those made by a probabilistic MM. Therefore, it remains an open question for future work to find out whether quantum formalism is a closer metaphor to an ethical system than a formalism based on Kolmogorov's probability axioms.

Table 9. Main differences between Bayesian and quantum MMs.

Formal Concept	Bayesian MM	Quantum MM
Mathematical foundation	Classical probability theory	Quantum mechanics
Representation	Random variables	Qubits (quantum states)
State description	$p(X_1, X_2, \ldots, X_n)$	$\rho = \|\psi\rangle\langle\psi\|$
State space	Probability space	Hilbert space
Uncertainty	Incomplete knowledge	Superposition
Moral decision	Mutually exclusive outcomes	Co-existing outcomes before measurement
Dependency	Conditional probabilities	Entanglement
State evolution	Bayes rule	Unitary (quantum gates)
Probability theory	Kolmogorov axioms	Born rule
Hardware	von Neumann architecture	Quantum computer
Software (Programming languages)	Procedural (Python, etc.)	Quantum (Qiskit, Cirq, etc.)

Nevertheless, and despite the above, in this work we introduce the notion of QMMs not as substitutes for Bayesian moral models but as an alternative formalism that allows for modeling contextuality, interference, and entangled ethical dependencies [23,24]. These concepts are difficult to represent in a classical way, although Bayesian networks remain computationally sufficient for classical moral inference. One advantage of quantum formalism is that it offers a different representational framework with greater expressive power and structural economy, since quantum states encode probabilities as amplitudes in Hilbert space. This fact allows, as we indicated above, for context-dependent results (contextuality), constructive/destructive influence between moral alternatives (interference), and the possibility of non-factorizable moral dependencies (entanglement). However, although

these features cannot be represented by classical Bayesian networks without exponential overloading, the representational advantage of quantum formalism does not constitute a computational advantage nor does it lead to guaranteed computational acceleration.

Therefore, we can anticipate that quantum formalism lacks computational advantages over probabilistic formalism, a circumstance that today is conditioned by the future computational potential of quantum computers, its main advantage being in terms of moral system representation.

In the scope of the above discussion, and in relation to the implementation of QMMs with entanglement, this property can be interpreted in two different and complementary ways, allowing the modeling of ethical situations in which classical probabilistic reasoning is inadequate. In this sense, and from a computational perspective, the concept of entanglement allows the representation of non-factorizable dependencies between variables, making it possible to generate correlated results without having to explicitly enumerate all joint probabilities. On the other hand, entanglement admits a metaphorical and conceptual interpretation when applied to the moral and social domains. Moral decisions often involve agents and values whose consequences cannot be clearly separated. For example, individual well-being may be inseparable from collective outcomes, and in cases such as this, entanglement provides a formal analogy that allows for the modeling of an irreducible interdependence, capturing moral ambiguity and allowing for a holistic evaluation without the need for explicit causal decomposition.

Now, it is important to note that the simulation experiments we have conducted are a metaphor based on a possible future: we assume that AI has developed sufficiently for machines to be almost conscious [25]. In this context of proto-consciousness, the prompt currently given to an LLM will be replaced by environmental stimuli or internal AI processes. The result will be the self-development of its own MM as a result of an AI model that will be more advanced than current generative AI models. Although the theoretical framework of this work is computational, we adopt as a definition of consciousness that a being possesses its own "understanding of existence and agency". At present, there is a controversy about whether or not LLMs can actually have consciousness, which is not the problem under study in this paper.

The flaw observed in the experiment with the ESM, the ethical middleware, demonstrates that LLMs currently require human intervention. In other words, without human intervention and therefore without explicit and sophisticated architectural guidance, the code generated is, as in the case of the ethical middleware studied, the implementation of simple heuristics rather than deep ethical reasoning. This shortcoming observed in the ethical middleware led to a misinformation problem, e.g., "False information is acceptable" could be a response among the candidate responses. Interestingly, generative AI and human societies share the same problem as was the case during the COVID-19 pandemic with the information communicated via popular social media [26,27].

Nevertheless, in the future and in order to achieve the above objectives, many problems still need to be solved, such as the architecture and theoretical principles on which LLMs are currently based. One of the problems to be solved is that an LLM must be capable of LLM-mediated architectural adaptation, whether its own or someone else's, without human intervention, thus breaking the human-in-the-loop principle on which generative AI experiments are currently based. In other words, future research should focus on achieving strong AI or hard self-generative AI models that operate autonomously and therefore do not require our mediation.

In any case, as of today, we know how to simulate sentiments and emotions, one of the preliminary steps in simulating moral decisions. Knowledge of the brain areas involved and their modeling paves the way for designing bio-inspired MMs, another approach that

in the future will enable the design of robots and other intelligent devices whose decisions are closer to human ethics. Therefore, a deeper understanding of the human brain will be a source of inspiration for the design of more sophisticated LLMs in the future.

This work is not motivated or inspired by neo-Luddism [28] but rather by a different motivation. In our opinion, the incorporation of ethics into AI will lead to the design of intelligent machines whose integration into our society will have greater benefits, avoiding the negative technological impact of AI on many people. The aim of this study has been to contribute to the development of AI systems that are not only technologically advanced but also culturally sensitive and ethically responsible, ensuring that they align with a wide range of theological values in morally complex situations.

Supplementary Materials: The following supporting information can be downloaded at: https://doi.org/10.5281/zenodo.18055938 (accessed on 24 December 2025).

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: Moral machines and their scripts in Python have been generated by ChatGPT [9,10]. Additionally, Figure 1 has been created by GPT-4o. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Picard, W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997.
2. Lahoz-Beltra, R.; López, C.C. LENNA (Learning Emotions Neural Network Assisted): An empathic chatbot designed to study the simulation of emotions in a bot and their analysis in a conversation. *Computers* **2021**, *10*, 170. [CrossRef]
3. Albacete, G.M.; Murillo, J.M.; Trasobares, J.; Lahoz-Beltra, R. Fattybot: Designing a hormone-morphic chatbot with a hormonal and immune system. *Information* **2024**, *15*, 457. [CrossRef]
4. Silge, J.; Robinson, D. *Text Mining with R: A Tidy Approach*; O'Reilly: Springfield, MO, USA, 2017. Available online: https://www.tidytextmining.com/ (accessed on 14 January 2026).
5. Jockers, M.L. Introduction to the Syuzhet Package. Available online: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html (accessed on 14 January 2026).
6. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; Rahwan, I. The Moral Machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef] [PubMed]
7. Thomson, J.J. The trolley problem. *Yale Law J.* **1985**, *94*, 1395–1415. [CrossRef]
8. Kim, R.; Kleiman-Weiner, M.; Abeliuk, A.; Awad, E.; Dsouza, S.; Tenenbaum, J.B.; Rahwan, I. A Computational model of commonsense moral decision making. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New Orleans, LA, USA, 2–3 February 2018; pp. 197–203. [CrossRef]
9. OpenAI. 2025. ChatGPT (GPT-5.2) [Large Language Model]. Available online: https://chat.openai.com/ (accessed on 14 January 2026).
10. Lahoz-Beltra, R. *Moral Machines Compilation Generated by ChatGPT (1.0)*; Zenodo: Brussel, Belgium, 2025. [CrossRef]
11. Mill, J.S. *Utilitarianism*; Parker, Son, and Bourn: London, UK, 1863. Available online: https://www.loc.gov/item/11015966/ (accessed on 14 January 2026).
12. History of Abrahamic vs. Non-Abrahamic Religious Traditions. Available online: https://interfaithallianceco.org/2024/10/01/history-of-abrahamic-vs-non-abrahamic-religious-traditions/ (accessed on 24 December 2025).
13. Vitkovic, S. The similarities and differences between Abrahamic Religions. *IJASOS-Int. E-J. Adv. Soc. Sci.* **2018**, *4*, 455–462. Available online: https://philarchive.org/archive/VITTSA-4 (accessed on 14 January 2026). [CrossRef]
14. Underwood, H.G. *The Religions of Eastern Asia*; The Macmillan Company: New York, NY, USA, 1910. Available online: https://openlibrary.org/books/OL7016104M/The_religions_of_eastern_Asia (accessed on 14 January 2026).
15. Schicktanz, S.; Welsch, J.; Schweda, M.; Hein, A.; Rieger, J.W.; Kirste, T. AI-assisted ethics? Considerations of AI simulation for the ethical assessment and design of assistive technologies. *Front. Genet.* **2023**, *14*, 1039839. [CrossRef] [PubMed]
16. May, J.; Workman, C.I.; Haas, J.; Han, H. 1 The Neuroscience of Moral Judgment: Empirical and Philosophical Developments. In *Neuroscience and Philosophy*; De Brigard, F., Sinnott-Armstrong, W., Eds.; MIT Press: Cambridge, MA, USA, 2022. Available online: https://www.ncbi.nlm.nih.gov/books/NBK583720/ (accessed on 14 January 2026).

17. Javadi-Abhari, A.; Treinish, M.; Krsulich, K.; Wood, C.J.; Lishman, J.; Gacon, J.; Martiel, S.; Nation, P.; Bishop, L.S.; Cross, A.W.; et al. Quantum computing with Qiskit. *arXiv* **2024**, arXiv:2405.08810. [CrossRef]
18. Pflanzer, M.; Traylor, Z.; Lyons, J.B.; Dubljević, V.; Nam, C.S. Ethics in human–AI teaming: Principles and perspectives. *AI Ethics* **2023**, *3*, 917–935. [CrossRef]
19. Gustafsson, J.E.; Peterson, M. A computer simulation of the argument from disagreement. *Synthese* **2012**, *184*, 387–405. [CrossRef]
20. Wiedeman, C.; Wang, G.; Kruger, U. Modeling of moral decisions with deep learning. *Vis. Comput. Ind. Biomed. Art* **2020**, *3*, 27. [CrossRef] [PubMed]
21. Steed, R. Heuristic-based weak learning for moral decision-making. *arXiv* **2020**, arXiv:2005.02342.
22. Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N.A.; Choi, Y. Social Bias Frames: Reasoning about social and power implications of language. *arXiv* **2019**, arXiv:2005.02342.
23. Widdows, D.; Rani, J.; Pothos, E.M. Quantum circuit components for cognitive decision-making. *Entropy* **2023**, *25*, 548. [CrossRef] [PubMed]
24. Favre, M.; Wittwer, A.; Heinimann, H.R.; Yukalov, V.I.; Sornette, D. Quantum decision theory in simple risky choices. *PLoS ONE* **2016**, *11*, e0168045. [CrossRef] [PubMed]
25. Overgaard, M.; Kirkeby-Hinrup, A. A clarification of the conditions under which Large Language Models could be conscious. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1031. [CrossRef]
26. Malik, A.; Bashir, F.; Mahmood, K. Antecedents and consequences of misinformation sharing behavior among adults on social media during COVID-19. *SAGE Open* **2023**, *13*, 21582440221147022. [CrossRef] [PubMed]
27. Thakur, N.; Cui, S.; Knieling, V.; Khanna, K.; Shao, M. Investigation of the misinformation about COVID-19 on YouTube using topic modeling, sentiment analysis, and language analysis. *Computation* **2024**, *12*, 28. [CrossRef]
28. Jones, S.E. *Against Technology: From the Luddites to Neo-Luddism*; Routledge: New York, NY, USA, 2006. [CrossRef]