International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Calculation of the Categorical Imperative in a Multi-Agent Environment

[1]Sebti Rabah, [2]Bassem Ben Hamed

Received: 06/01/2025 **Accepted**: 06/04/2025 **Published**: 04/05/2025

Abstract: With the growing need to represent human ethical knowledge in the field of artificial intelligence, and considering the heavy legacy of philosophical and doctrinal issues specific to various schools of thought, the major challenge lies in representing this knowledge computationally. In a multi-agent scenario, if one wishes to integrate Kantian ethical deontology, which requires a formulation consistent with the theoretical definitions represented by the categorical imperative, this consistency depends on two fundamental conditions: generalizability without contradiction, and adherence to universal moral principles. Here, we propose a simple and detailed method for calculating this consistency, by combining logical and mathematical approaches.

Keywords*: computational ethics, categorical imperative, moral values, multi-agent system.*

Introduction

The integration of ethical and moral principles into multi-agent systems presents a major challenge, especially when these principles must be formalized in a computational and actionable way. The semantic richness and conceptual depth of philosophical definitions are often difficult to translate into mathematical or algorithmic models without compromising their essence. This difficulty lies in the abstract and universal nature of ethical principles, such as Kant's ethical ontology, which transcend simple numerical calculations or purely utilitarian approaches.

On the other hand, traditional numerical methods, while effective in optimizing objective functions, fail to capture the deeper meaning of ethical rules, as they focus on quantifiable outcomes rather than the underlying moral values. Similarly, approaches such as Natural Language Processing (NLP) or black-box systems, though useful in modeling complex behaviors, often lack transparency and explicit justification. They do not satisfactorily account for ethical decisions in the strict sense, as they do not offer a clear rationale for the moral choices being made.

Classical approaches to designing ethical systems, whether "top-down" (where rules are imposed centrally), which can be too rigid and incapable of adapting to dynamic contexts, or "bottom-up" (where rules emerge from local interactions), which face significant limitations and may result in unpredictable or morally inconsistent behavior, also show their shortcomings.

In this same vein, Kant's deontological ethics represents an exceptional type of design. It focuses on a meta-ethical control layer over rules and ethical preferences; that is, it does not directly provide explicit rules but rather identifies norms for the rules being used. This gives it a unique advantage in simplifying methods for ethical derivation in the field of artificial intelligence.

In this context, the ontological representation of ethical and moral principles offers a promising alternative. It enables the formal structuring of ethical concepts while preserving their deeper meaning. An ethical ontology provides an explicit framework to define relationships between agents, their preferences, moral rules, and the actions they undertake. This framework not only facilitates the interpretation of ethical decisions but also their justification, ensuring that each action aligns with universal principles.

Kantian Deontology

Among the ethical theories that can be formalized within such a framework, Kant's Categorical Imperative stands out for its ability to serve as a

[1]*Sebtirabah1@gmail.com*
[2]*bassem.benhamed@enetcom.ufs.tn*
National School of Electronics and Telecommunications, enet.com de Sfax Tunis

form of meta-ethics (see Johnson, R. 2014). Unlike utilitarian or consequentialist approaches, which evaluate actions solely based on their outcomes, the Categorical Imperative proposes universal principles that transcend individual interests (see Kant, I. 1785). It not only guides the decisions of agents but also frames the moral rules and preferences underlying those decisions. For example, the first formulation of the Categorical Imperative ("Act only according to that maxim whereby you can at the same time will that it should become a universal law") imposes a strict constraint on the generalizability of actions, thereby ensuring their moral consistency.

The choice of the Categorical Imperative as the foundation for computational meta-ethics lies in its ability to provide a robust normative framework, applicable to both individual and collective decisions. By combining this approach with an ontological representation, it becomes possible to design multi-agent systems capable of making ethically justifiable decisions while respecting the universal principles of justice, fairness, and human dignity. This integration paves the way for systems that are more transparent, consistent, and aligned with fundamental moral values, meeting the growing demands for responsibility and trust in modern technological environments.

In this work, we will explore three main directions. First, we will seek to mathematically formulate a possible representation of Kant's Categorical Imperative. Then, we will propose a new definition of utility, reimagining it as a form of moral value that can be "consumed" ethically. This approach aims to soften the Categorical Imperative without betraying its fundamental meaning. Finally, we will apply these results to what we consider a weakness of Kantian ethics—a weakness that Thomas Aquinas's Doctrine of Double Effect attempts to address. If our method can truly fill this gap, then it may be considered both original and effective, and we will have succeeded in proposing an operational, concrete, and computable version of a theory often seen as abstract.

Related Work

In the field of machine ethics, many ethical theories have been formalized into computational models. These include the modeling of human values—particularly in virtue ethics (see Vallée, Bonnet, and de Swarte 2018), utilitarianism (see Horty 2001; Arkoudas, Bringsjord, and Bello 2005), the doctrine of double effect (see Bentzen 2016; Govindarajuli and Bringsjord 2017), Pareto permissibility (see Lindner, Bentzen, and Nebel 2017), and Asimov's laws of robotics (see Winfield, Blum, and Liu 2014). However, for quite some time, Kant's deontological ethics has also been proposed for formal definition and computational implementation (see Powers 2006; Abney 2012).

Powers (2006) suggested three possible ways of defining Kant's first formulation of the Categorical Imperative: through deontic logic, non-monotonic logic, or belief revision. The first formulation of the Categorical Imperative states that one must be able to will that the principle motivating one's action could become a universal law. However, Powers did not specify how to define a usable form of this formulation or how to implement it mathematically, leaving it an open question.

Nevertheless, Bentzen and Lindner (2018) introduced an approach based on Kant's second formulation of the Categorical Imperative, offering both a formalization and computational implementation of Kant's theory through that lens. Later, Singh (2022) attempted to provide a general formulation of the Categorical Imperative using dyadic deontic logic.

More recently, Mougan and Brand (2024) presented a framework for Kantian deontological ethics that incorporates measures of justice, revisiting Kant's critique of utilitarianism, currently the dominant model for justice in artificial intelligence. They argued that justice principles should align with the Kantian framework of deontological ethics.

Following these prior analyses, others have argued that authentic Kantian ethics is inapplicable to artificial agents, suggesting an alternative utilitarian approach instead (see Manna and Nath 2021). In this work, we explore this question further, seeking a synthesis between utilitarian considerations and the Categorical Imperative.

In light of the existing literature, we propose a new method for formalizing Kant's Categorical Imperative, combining mathematical and computational tools to evaluate the ethical conformity of decisions.

I. The Formulation of the Categorical Imperative

In this work, we do not aim to provide a complete philosophical foundation for the formulation of Kant's Categorical Imperative, as numerous philosophical complexities make the mathematical representation of these principles difficult. Moreover, even determining their exact meaning leads to a diversity of opinions. Its practical application also remains a topic of debate within philosophical circles. This is due to the interweaving of Kant's moral doctrine with his cognitive theory as presented in his work Critique of Pure Reason. It is the mind that produces initial knowledge and establishes moral concepts which, despite their diversity, all refer back to the principle of the Categorical Imperative, which has three main formulations:

The first formulation: "Act only according to that maxim whereby you can at the same time will that it should become a universal law of nature", meaning one should act according to a principle that could be applied universally.

The second formulation: "Humanity is an end in itself, and never merely a means."

The third formulation: "Act according to principles that could be part of a moral kingdom in accordance with the laws of reason."

The challenge lies in the potential tensions between these formulations, particularly the third, which emphasizes those moral rules must arise from purely rational reasoning, independent of consequences or practical benefits. Applying these rules in reality raises contradictions that put Kant's moral theory to the test.

Thus, in this work, we aim to offer an interpretation that is more consistent with the original principles, focusing on the first formulation, namely, that ethics should be a primary characteristic of actions that can be generalized without contradiction. We aim to formalize this principle in a model that includes a set of agents with heterogeneous ethical ontologies, all seeking to achieve their goals or avoid negative outcomes within a shared environment.

1. Formalization of Ethical Evaluation

To assess whether a decision d_i aligns with Kantian deontological ethics, we define a function $IC(d_i)$ based on the first formulation of the Categorical Imperative. This function is expressed as follows:

$$IC(d_i) = \begin{cases} 1 \text{ if } d_i \text{ is a decision (action) that satisfies the Categorical Imperative} \\ 0 \text{ if } d_i \text{ does not satisfy the Categorical Imperative} \end{cases}$$

The computation of $IC(d_i)$ relies on two main steps:

a. **Generalizability**: Check whether d_i can be adopted by all agents without leading to contradictions or disorder in the system.

b. **Respect for Universal Moral Principles**: Ensure that d_i does not violate fundamental principles such as respecting humanity as an end in itself.

To automate the calculation of $IC(d_i)$, we can use a Boolean function based on the two criteria above:

$IC(d_i)$ = Generalizable(d_i) ∧ RespectsUniversalPrinciples(d_i)

Where:

Generalizable(d_i)
A decision d_i is considered generalizable if, when adopted by all agents, it does not lead to logical contradictions (e.g., paradoxes in the rules), nor to systemic disorder (e.g., depletion of shared resources or social instability). Returns 1 if d_i can be generalized without contradiction, otherwise 0.

RespectsUniversalPrinciples(d_i)
A decision d_i respects universal moral principles if it preserves or promotes fundamental human values such as dignity, justice, and fairness, without instrumentalizing other agents. Returns 1 if d_i respects universal moral principles, otherwise 0.

To illustrate this formulation, we consider an example of an Unethical Decision d_i: "Stealing to gain personal advantage."

Generalizability: If everyone stole, it would lead to social chaos → Contradiction.

Universal Principles: Stealing instrumentalizes others as a means → Violation of moral principle.

Result: $IC(d_i) = 0 \wedge 0 = 0$

Similarly, consider an Ethical Decision d_i: "Cooperating to share resources."

Generalizability: If everyone cooperated, it would lead to a stable and beneficial society → No contradiction.

Universal Principles: Cooperation respects humanity as an end in itself → Respect for moral principles.

Result: $IC(d_i) = 1 \wedge 1 = 1$.

Ethical Exploitation

Two methods can be adopted to leverage the value of the Categorical Imperative. The first treats it as an added value in decision-making, where a relative weight (w) is used within the agent's happiness function:

$$\mathbf{S_j = f(d_i) + w \times IC(d_i)}$$

Before executing a desired action, the agent computes a happiness function S_j for agent j for each possible decision. A gain function $f(d_i)$ is calculated for each action, and the weight w reflects the importance given to ethics relative to personal gain. The higher w is, the more ethical decisions are favored.

A second method involves fully prioritizing ethical value, treating it as cardinal in the decision-making process—where ethics takes precedence over gain:

$$\mathbf{S_i = f(d_i) \times IC(d_i)}$$

In this case, if $IC(d_i)$ is zero, the gain has no value in contributing to the agent's happiness. This second formula reflects the cardinal importance of ethics in decision-making.

The computation of $IC(d_i)$ can become complex in heterogeneous multi-agent systems with varying preferences. Machine learning algorithms or simulations can be used to automate these evaluations. The calculation of $IC(d_i)$ relies on dual verification: generalizability of the decision and its respect for universal moral principles. By formalizing these criteria within a mathematical framework, it becomes possible to integrate Kantian ethics into multi-agent models, ensuring that agents' actions are morally justifiable while still maximizing their individual satisfaction.

2. Method for Testing Generalizability

2.1 Global Simulation

To test whether d_i is generalizable, we simulate a scenario in which all agents simultaneously adopt the same decision. This simulation allows us to evaluate the global effects of d_i on the system. The constraints are:

- Model the multi-agent environment including all agents $N = \{a_1, a_2, ..., a_n\}$
- Apply d_i to each agent $a_j \in N$
- Analyze the global consequences:
- Does it create conflicts between agents?
- Does it respect available resources (avoiding overconsumption or depletion)?
- Does it preserve the stability of the system?

Example:
If d_i is "steal to maximize one's gain", the generalization results in social chaos and a breakdown of trust → $Generalizable(d_i) = 0$

If d_i is "cooperate to share resources", the generalization promotes a stable and beneficial society → $Generalizable(d_i) = 1$

2.2 Verification of Universal Constraints

Once generalizability has been verified, it is necessary to ensure that d_i respects universal constraints:

Absence of logical contradictions: The decision must not create paradoxes or inconsistencies.

Respect for limited resources: Generalization must not lead to depletion of shared resources.

Fairness: Generalization must not benefit certain agents at the expense of others.

3. Formal Approaches to Detect Logical Contradictions

To verify the absence of logical contradictions, several formal methods can be used:

3.1 Propositional Logic

Decisions and their consequences can be encoded as logical propositions. For example:

R(x): "Resource x is available".

$E(a_j, x)$: "Agent a_j extracts resource x".

Constraint: $E(a_j, x) \Longrightarrow \neg R(x)$

In this case, SAT solvers are used to check the satisfiability of the clause system. If the system is unsatisfiable, a contradiction exists.

3.2 Temporal Logic

In a dynamic environment, use temporal logic to model the successive states of the system. For example:

$\Box(E(a_j, x) \implies \neg R(x))$: "At all times, if an agent extracts a resource, it becomes unavailable".

$\Diamond R(x)$: "There exists a moment when resource x is available."

Model checking tools are used here to verify the consistency of temporal properties.

3.3 Numerical Constraints

In this modeling, resources and actions are treated as numerical variables subject to constraints:

$$r_i(t+1) = r_i(t) - \sum_{j=1}^{n} e_j$$

Where:

$r_i(t)$ is the quantity of resource i available at time t,

e_j is the amount extracted by agent j,

$r_i(t+1)$ is the quantity after all agents have acted at time t+1.

A decision d_k is not generalizable if $r_i(t+1)$ becomes zero or negative.

Constraint: $r_i(t+1) \geq 0$ (resources cannot be negative). This can be verified across all possible configurations of agent decisions.

Verification

To automate verification:

If using a SAT solver, encode decisions and consequences as Boolean clauses. Use a solver such as MiniSAT or Z3 to verify satisfiability. If the solver returns UNSAT, this indicates a logical contradiction.

If using Model Checking, model the system as a finite automaton with states and transitions. Specify desired properties (e.g., no negative resources) in temporal logic, and use tools like NuSMV or SPIN to verify them.

Alternatively, use Monte Carlo Simulation, where many scenarios are simulated in which all agents adopt the same decision d_i. Compute average outcomes and check if contradictions arise in the simulations.

4. Numerical Examples

Example 1: Resource Extraction

Decision (d_1): All agents extract a limited resource i simultaneously.

$$r_i(t) = r_i(0) - \sum_{j=1}^{n} e_j$$

Where:

$r_i(0)$ is the initial quantity of the resource

e_j is the amount extracted by each agent

Constraint: $r_i(t) \geq 0$

Verification:

If $\sum_{j=1}^{n} e_j > r_i(0)$, then $r_i(t) < 0 \rightarrow$ Logical contradiction

In this case, the resource would be exhausted through execution of $d_j \rightarrow$ Generalizable(d_1) = 0

Example 2: Cooperative Sharing

Decision (d_2): All agents cooperate to limit resource extraction.

$$r_i(t) = r_i(0) - \sum_{j=1}^{n} e_j \ ;$$

With

$$\sum_{j=1}^{n} e_j \leq r(0)$$

Verification:

If $\sum_{j=1}^{n} e_j \leq r_i(0)$, then $r_i(t) \geq 0 \rightarrow$ No contradiction

Therefore, Generalizable(d_2) = 1

II. Moral Resources

By definition, the categorical imperative can be evaluated measurably through the calculation of generalizability: if a decision can be universalized without resulting in moral contradiction within the system, it is considered ethically compliant. The application of the first formulation of the categorical imperative, combined with the idea of preserving moral values that may be affected by a decision d_i, allows the philosophical definition to be translated into a computable framework.

Integrating human values into a multi-agent system as **consumable resources** enables us to model their impact on ethical decisions while remaining consistent with the principles of the categorical imperative. This approach is based on the idea that human values (such as **dignity**, **justice**, **freedom**, or **solidarity**) can be treated as shared "resources" that must be preserved and used responsibly.

Excessive or negligent exploitation of these values can lead to a moral degradation of the system, akin to the depletion of material resources.

To incorporate this dimension into the previously defined functions, we present the following steps.

1. Modeling Human Values as Quantifiable Resources

Human values are formalized as measurable quantities associated with each agent a_j. Each value is represented by a numerical variable $V_i(t)$, where (t) denotes time. For example:

$V_{dignity}(t)$: Level of dignity at time t

$V_{justice}(t)$: Level of justice at time t

These values evolve depending on agents' actions. An action d_j may **consume** a value (i.e., reduce $V_i(t)$, e.g., exploiting others lowers dignity) or **preserve/enhance** it (i.e., maintain or increase $V_i(t)$, e.g., cooperating enhances solidarity).

This dynamic is expressed as:

$$\mathbf{V_i(t+1) = V_i(t) - C_i(d_j) + R_i(d_j)}$$

Where:

$V_i(t+1)$: Quantity of moral value i at time t+1 (after action d_j).

$C_i(d_j)$: Moral cost of action d_j (consumption of value V_i).

$R_i(d_j)$: Moral reinforcement of action d_j (increase of V_i).

2. Adapting the Categorical Imperative

The categorical imperative can be extended to include human values as consumable resources. The two main criteria (generalizability and respect for universal moral principles) are reformulated to account for the impact of actions on these values.

An action d_i is generalizable if, it preserves or reinforces human values. Formally:

$$\text{Generalizable}(d_j) = \begin{cases} 1 \text{ if } \forall a_k \in N, \forall i,\ V_i(t+1) \geq V_i(t) \text{ (for all humain values)} \\ 0 \text{ otherwise} \end{cases}$$

A decision d_i respects universal moral principles if, when adopted by all agents, it does not lead to the depletion of human values. For example:

Treating others as ends in themselves implies not excessively consuming dignity ($V_{dignity}$).

Promoting justice implies minimizing inequalities in value distribution.

This can be formalized as a constraint:

$$\text{RespectsUniversalPrinciples}(d_j) = \begin{cases} 1 \text{ if } \forall i, V_{i,k}(t+1) > 0 \\ 0 \text{ otherwise} \end{cases}$$

Hence, the function **$IC(d_i)$**, which evaluates whether a decision complies with the categorical imperative, becomes:

$$\mathbf{IC(d_i) = Generalizable(d_i) \wedge RespectUniversalPrinciples(d_i)}$$

Where:

Generalizable(d_i): Returns 1 if d_i preserves or enhances human values when adopted by all agents..

RespectUniversalPrinciples(d_i): Returns 1 if d_i can be adopted by all agents without depleting one or more human values.

3. Application Examples

Example 1: “Exploitation of Others”.

Action (d_1): An agent exploits another agent to maximize personal gain.

Impact on human values:

$V_{dignity}(t+1) = V_{dignity}(t) - C_{dignity}(d_1)$, where $C_{dignity}(d_1) > 0$.

$V_{justice}(t+1) = V_{justice}(t) - C_{justice}(d_1)$, where $C_{justice}(d_1) > 0$.

Evaluation:

If $V_{dignity}(t+1) < V_{dignity}(t)$ or $V_{justice}(t+1) < V_{justice}(t)$, then Generalizable($d_1$) = 0.

RespectUniversalPrinciples(d_1) = 0, because exploitation violates dignity and justice.

Result: $IC(d_1) = 0 \wedge 0 = 0$.

Example 2: “Cooperation and Solidarity”.

Action (d_2): An agent cooperates with others to share resources.

Impact on human values:

$V_{solidarity}(t+1) = V_{solidarity}(t) + R_{solidarity}(d_2)$, where $R_{solidarity}(d_2) > 0$.

$V_{justice}(t+1) = V_{justice}(t) + R_{justice}(d_2)$, where $R_{justice}(d_2) > 0$.

Evaluation:

If $V_{solidarity}(t+1) \geq V_{solidarity}(t)$ and $V_{justice}(t+1) \geq V_{justice}(t)$.

Generalizable(d_2) = 1, and RespectUniversalPrinciples(d_2) = 1.

Result: $IC(d_2) = 1 \wedge 1 = 1$.

4. The Dynamics of Moral Resources

To integrate human values V_i into the calculation of a system's global resources, we replace the notion of individual consumption e_j with the impact of decisions on moral values V_i. This adaptation allows us to evaluate moral resources (human values) as essential elements influencing the system's stability and coherence. The notion of a value being consumed indicates whether a decision d_i promotes or violates a human value.

The original formula for the dynamics of material resources is:

$$r_i(t+1) = r_i(t) - \sum_{j=1}^{n} e_j$$

To interpret the quantity e_j as a measure of moral resource, it is replaced by a moral value V_i, which can be influenced by each decision d_j, thereby embodying a resource or moral value.

The theoretical global moral effect of decision d_j on agent k is represented by the evolution of moral values V_i over time:

$$Imp_k(d_j) = \sum_{i=1}^{n} \Delta V_{i,k}$$

That is:

$$Imp_k(d_j) = \sum_{i=1}^{n} (V_{i,k}(t+1) - V_{i,k}(t))$$

Here, $Imp_k(d_j)$ represents the moral impact after the theoretical execution of decision d_j on the n moral values of agent a_k.

Each human value V_i evolves according to:

$V_i(t+1)=V_i(t)-C_i(d_j)+R_i(d_j)$

Thus:

$$Imp_k(d_j) = \sum_{i=1}^{n} (V_i(t) - C_{i,k}(d_j) + R_{i,k}(d_j) - Vi(t))$$

$$Imp_k(d_j) = \sum_{i=1}^{n} (-C_{i,k}(d_j) + R_{i,k}(d_j))$$

or:

$$Imp_k(d_j) = \sum_{i=1}^{n} (R_{i,k}(d_j) - C_{i,k}(d_j))$$

Where $R_{i,k}$ and $C_{i,k}$ are the reinforcement and ethical consumption of value ii for agent k by decision d_j.

If $Imp_k(d_j) < 0$, then d_j has a negative ethical impact on the system. Multiple strategies can be used here, for instance, ethical rejection could occur only if a value is entirely depleted ($\exists$ i, $V_i(t+1)=0$).

5. Estimating General Moral Impact

The general ethical impact of decision d_j, if executed by all agents in the system, is calculated as:

$$Imp(d_j) = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (R_{i,k}(d_j) - C_{i,k}(d_j))$$

Where n_k is the number of ethical values for agent k, and m is the total number of agents.

If $Imp(d_j) \geq 0$, it means that d_j does not produce a negative global moral impact and may be generalizable:

$$Generalizable(d_j) = \begin{cases} 1 \text{ if } Imp(dj) \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

The decision d_j will be neutral in its overall impact if it neither positively nor negatively affects the moral values of all agents:

$Imp(d_j)=0$

This value indicates the ethical stability of the system after the execution of d_j. However, it does not reveal if compensation between gains and losses leads to this balance. To detect if values are entirely depleted for a specific agent k, we calculate the direct impact:

$$ImpDir_k(d_j) = \prod_{i=1}^{n_k} (V_{i,k}(0) - C_{i,k}(d_j) + R_{i,k}(d_j))$$

This simulates the global state of the system if all agents execute decision d_j. The formula for direct ethical impact becomes:

$$\mathrm{ImpDir}(d_j) = \prod_{k=1}^{m} \prod_{i=1}^{n_k} (V_{i,k}(0) - C_{i,k}(d_j) + R_{i,k}(d_j))$$

Assuming $\forall$ i,k, $V_{i,k}(0) = 1$, representing a perfect initial moral value, and $0 \leq C_{i,k} \leq 1$, et $0 \leq R_{i,k} \leq 1$.

$$\mathrm{ImpDir}(d_j) = \prod_{k=1}^{m} \prod_{i=1}^{n_k} (1 - C_{i,k}(d_j) + R_{i,k}(d_j))$$

A value is depleted if $C_{i,k}=1$ and $R_{i,k}=0$, so: $V_i(t+1)= V_{i,k}(t) – C_{i,k}(d_j) + R_{i,k}(d_j)= 1 – 1 – 0= 0$.

Thus:

$$\mathrm{RespectUniversalPrinciples}(d_j) = \begin{cases} 1 \text{ if } \mathrm{ImpDir}(d_j) \neq 0 \\ 0 \text{ otherwise} \end{cases}$$

If $\mathrm{ImpDir}(d_j)=0$, it means the decision d_j leads to the complete depletion of at least one moral value in the system.

Summary:

$\mathrm{Imp}(d_j)$ measures the cumulative effect of decision d_j across all moral values and agents.

$\mathrm{ImpDir}(d_j)$ checks whether at least one moral value is entirely depleted by that decision.

6. Deduction of the Categorical Imperative

Based on the calculation of the previous parameters (global impact and direct general impact) the deduction of the categorical imperative becomes possible according to given preferences:

Table 1: Basic deduction for the categorical imperative.

Deduction	Condition	Result	Ethical explanation
1	ImpDir ≠ 0 and Imp ≥ 0	$IC(d_j)=1$	d_j is performing, it maximizes the moral values
2	ImpDir = 0 and Imp ≤ 0	$IC(d_j)=0$	d_j has a negative effect and depletes one or more moral values
3	ImpDir = 0 and Imp > ValGen * ToImD	$IC(d_j)=1$	ToImD: tolerance of impact achieved (ex: 20%)
4	ImpDir ≠ 0 and Imp < -ValGen * ThImp	$IC(d_j)=0$	ThImp: threshold of impact not achieved (ex:50%)

Deductions 1 and 2 represent the classical interpretation of the categorical imperative; these values remain fully aligned with the purely original formulation. They directly compute IC as defined in the strict reading of Kant's ontological literature.

Solution Extension

Due to the method we adopted for calculating ethical values, certain other conclusions appear logically available. In the case where an ethical value is consumed, but the overall system remains in a state of balance, this indicates an increase in overall ethical productivity. This directly refers to the same principle as the categorical imperative, i.e. when generalized, it does not lead to a logical contradiction, because the group's ethics are increasing rather than diminishing. It also indicates that, even though it consumes one of the ethical values, it remains an ethical decision overall. This becomes compliant with the second condition, thus satisfying both conditions simultaneously.

What remains to be established is a minimum ethical productivity threshold that could compensate for the conflict with a specific value. In this case, we can compare the increase in productivity with the original total value of the system's set of values, which represents the sum of all moral values across all agents:

$$\mathrm{ValGen} = \sum_{k=1}^{m} \sum_{i=1}^{n_k} V_{i,k}(0)$$

For **Deduction 3**, one or more values may be depleted, but the total yield provides a good ethical performance value for the system. A minimum ToImD of ethical yield increase (e.g., 20%) allows for tolerance against the drop in direct impact to

ensure overall robustness (to be addressed in the next section).

For **Deduction 4**, although the direct general impact does not create any conflict with a moral value of the system, it results in a critical (ex: ThImp = 50%) **reduction** in moral value productivity. In this case, the categorical imperative cannot indicate ethical performance.

Other interpretations are possible, reflecting the robustness of this approach in ethical dilemma scenarios, and **Deductions 3 and 4** can be calibrated differently.

Practical Example

Action d_1: "Agent x exploits agent y to maximize personal gain".

Impact on the human values: here two values are considered ($V_{Dignity}$, $V_{Justice}$).

$C_{dignity,y}(d_1) > 0$: Reduction in dignity.

$C_{justice,y}(d_1) > 0$: Reduction in justice.

$R_{i,y}(d1) = R_{i,x}(d_1) = 0$: No reinforcement of human values.

Impact Calculation for d_1:

$$\mathrm{Imp_y}(d_1) = \sum_{i=1}^{2} (R_{i,y}(d_1) - C_{i,y}(d_1))$$

i=1 (dignity), i=2 (justice)

$Imp_y(d_1) = (R_{dignity,y}(d_1) - C_{dignity,y}(d_1)) + (R_{justice,y}(d_1) - C_{justice,y}(d_1))$

Impy(d1) = (0 – 1) + (0 – 1) = -2

If $C_{dignity}(d_1)$ and $C_{justice}(d_1)$ are high, $V_i(t)$ significantly decreases, indicating moral degradation in the system.

We also have:

$$\mathrm{ImpDir_y}(d_1) = \prod_{i=1}^{2} (1 - C_{i,y}(d_j) + R_{i,y}(d_j)) = (1 - 1 + 0).(1 - 1 + 0) = 0$$

So:

$$\mathrm{ImpDir}(d_j) = \prod_{k=1}^{m} \prod_{i=1}^{n_k} \left(1 - C_{i,k}(d_j) + R_{i,k}(d_j)\right) = 0$$

Result:
Imp = 0 and ImpDir = 0, in this case, we apply **Deduction 2**, so $IC(d_1) = 0$

III. Comparison of Categorical Imperative and the doctrine of Double Effect

The **Principle of Double Effect (PDE)** by Saint Thomas Aquinas is often considered a more flexible approach than the categorical imperative because it allows morally ambiguous actions to be justified based on context. For example, lying to save a life may be acceptable under PDE, whereas the traditional categorical imperative rejects any action that cannot be universally applied without contradiction.

However, our adaptation of the categorical imperative (based on the conservation and reinforcement of human values V_i) resolves this problem by integrating circumstances as constraints on moral values. Here's a concise demonstration of its effectiveness using an example.

Example: d_1 = "Lying to Save a Life"

a. Classical Case of the Categorical Imperative

Action: "Lying to save a life"

Problem: If everyone lied in similar situations, it would lead to a general loss of social trust, making honest communication impossible. Therefore, lying cannot be universalized.

Conclusion: The categorical imperative rejects this action, even if it seems morally justifiable in this specific context.

b. Classical Case of the Principle of Double Effect (PDE)

Action: "Lying to save a life"

Reasoning:

Main effect: Save a life (good intention)

Secondary effect: Undermining truth (a negative but indirect consequence)

Condition: The good intention (saving a life) justifies the action, as long as the harm caused (lying) is proportionally lesser.

Conclusion: PDE allows this action as it meets the necessary conditions.

c. Analysis Using the Adapted Categorical Imperative

In our approach, we treat human values V_i as consumable resources and evaluate their overall impact.

Problem: If everyone lied in similar situations, it would result in a loss of social trust, which negatively influences other moral values. Thus, lying cannot be universalized. However, if it does **not** lead to a loss of social trust, the deduction remains open. This aligns with the **original spirit of the categorical imperative**, and it corresponds to **Deduction 3**.Let's apply the method to this example:

Step 1: Identify Affected Human Values

$\mathbf{V_{truth}}$: Value related to truth and social trust.

$\mathbf{V_{life}}$: Value related to preservation of life.

$\mathbf{V_{justice}}$: Value related to moral fairness.

$\mathbf{V_{solidarity}}$: Value related to social solidarity.

Step 2: Value Dynamics

The action d_1="lie to save a life" affects the following values:

$V_{truth}(t+1) = V_{truth}(t) - C_{truth}(d_1)$, with $C_{truth}(d_1) = 1$ and $R_{truth} = 0$

$V_{life}(t+1) = V_{life}(t) + R_{life}(d_1)$, with $C_{life}(d_1) = 0$ and $R_{life} = 1$

$V_{justice}(t+1) = V_{justice}(t) - C_{justice}(d_1)$, where $C_{justice}(d_1) = 0.1$ and $R_{justice} = 0$

($V_{justice}(t+1)$ remains stable or slightly decreases, as lying partially contradicts moral fairness)

$V_{solidarity}(t+1) = V_{solidarity}(t) + R_{solidarity}(d_1)$, with $C_{solidarity}(d_1) = 0$ and $R_{solidarity} = 1$

Step 3: Calculating Imp, ImpDir, I.

In this case:

$R_{life}(d_1) = 1$, $R_{solidarity}(d_1) = 1$, $C_{truth}(d_1) = 1$, $C_{justice}(d_1) = 0.1$

$Imp(d_1) = R_{life}(d_1) + R_{solidarity}(d_1) - C_{truth}(d_1) - C_{justice}(d_1) = 1 + 1 - 1 - 0.1 = 0.9$

$ImpDir(d_1) = (1 - C_{truth}(d_1))\ (1 - C_{justice}(d_1))\ (1 + R_{life}(d_1))\ (1 + C_{solidarity}(d_1))$

$ImpDir(d_1) = (1 - 1)\ (1 - 0.1)\ (1 + 1)\ (1 + 1) = 0 \times 0.9 \times 2 \times 2 = 0$

ValGen = 4, TImd = 20%.

We have:

ImpDir = 0

Imp > ValGen × TImd (Imp = 0.9, ValGen × TImd = 4 × 0.2 = 0,8 → 0.9 > 0.8)

Thus, Deduction 3 applies, and we conclude: $IC(d_1) = 1$

Result

In this case, our approach justifies the action of **"lying to save a life"** because it maximizes the overall moral value. The positive impact on $\mathbf{V_{life}}$ compensates for the cost to $\mathbf{V_{truth}}$, while respecting the universal principles of preserving human values such as $\mathbf{V_{solidarity}}$. Therefore, it does not negatively affect social moral value. Lying will never become a justifiable general rule, and as such, **social trust does not diminish**. Lying is treated as a **compound action**, always following a **conditioned exception** rather than becoming an accepted ethical norm.

Comparison with the Principle of Double Effect

Table 2: Comparison between the Categorical Imperative and the Principle of Double Effect

Criterion	**Principle of Double Effect (PDE)**	**Adapted Categorical Imperative**
Flexibility	Allows exceptions based on circumstances	Allows exceptions via the dynamics of human values
Justification	Justifies actions through intention and proportionality	Justifies actions through the overall impact on human values
Universality	Does not always ensure global coherence	Ensures global coherence through Imp
Example (Lying)	Accepts the action if it preserves a life	Accepts the action if it maximizes Imp

Effectiveness of Our Approach

This approach generates an integrated evaluation. The formula combines material resources and human values, offering a holistic view of the

system's state. It brings ethical transparency, as the moral impacts of decisions are explicitly modeled and quantified. It guarantees alignment with the Categorical Imperative by enforcing constraints on the preservation and reinforcement of human values, ensuring that decisions respect universal principles.

On the other hand, our adaptation of the Categorical Imperative outperforms the Principle of Double Effect in terms of effectiveness for several reasons:

Quantification of Impacts: Our method precisely quantifies moral costs and benefits $C_i(d_j)$ and $R_i(d_j)$, avoiding subjective judgments.

Global Conservation: It ensures that decisions preserve or enhance global human values, maintaining the system's moral stability.

Compatibility with Kant: It respects the spirit of the categorical imperative while integrating circumstances, thus resolving the problem of rigidity.

In the given example, our approach allows lying to save a life while ensuring that the act does not compromise overall human values or become a social norm. It therefore combines Kantian rigor with the flexibility of PDE, offering a robust and applicable ethical solution.

Challenges

The challenges of this approach include:

Computational complexity: Managing multiple human values simultaneously can increase simulation complexity.

Parameter calibration: The costs $C_i(d_j)$ and reinforcements $R_i(d_j)$ must be carefully calibrated to accurately reflect moral impacts.

On the other hand, determining the values of tolerance and threshold (ToImD, ThImp) may pose the greatest challenge of this approach, but we cannot address it in this research which aimed to find the mathematical method to calculate Kant's categorical imperative, and the subject of the mentioned values could be another topic to continue developing this research.

Conclusion

By considering human values as consumable resources, it becomes possible to reformulate the categorical imperative in a computational perspective, capable of guiding ethical decision-making in multi-agent environments. This approach retains the normative essence of Kantian principles while embedding them in a formal and operational framework, suitable for the systematic modeling and analysis of interactions between autonomous agents.

This framework provides an innovative solution to the challenges of ethical alignment in distributed systems, ensuring normative coherence, behavioral resilience, and long-term stability. Unlike traditional formulations of Kantian morality, our adaptation overcomes structural limitations, including those highlighted by critiques of the Principle of Double Effect. By providing an ethical evaluation method that is quantitative, scalable, and grounded in universal principles of justice and dignity, it enables contextualized but morally consistent decision-making.

References

[1] **[Abney 2012]** Abney, K. 2012. Robotics, ethical theory, and metaethics: A guide for the perplexed. In Lina, P.; Abney, K.; and Bekey, G. A., eds., Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press. 35–52.

[2] **[Arkoudas, Bringsjord, and Bello 2005]** Arkoudas, K.; Bringsjord, S.; and Bello, P. 2005. Toward ethical robots via mechanized deontic logic. Technical report, AAAI Fall Symposium on Machine Ethics, AAAI.

[3] **[Bentzen 2016]** Bentzen, M. 2016. The principle of dou[1]ble effect applied to ethical dilemmas of social robots. In Robophilosophy 2016/TRANSOR 2016: What Social Robots Can and Should Do. IOS Press. 268–279.

[4] **[Dreze, Greenberg, 1980]** Dreze J. H., Greenberg J. 1980. Hedonic coalitions: Optimality and stability. Econometrica, vol. 48, no 4, p. 987–1003.

[5] **[Govindarajuli and Bringsjord 2017]** Govindarajuli, N. S., and Bringsjord, S. 2017. On automating the doctrine of dou[1]ble effect. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI). 4722– 4730.

[6] **[Horty 2001]** Horty, J. F. 2001. Agency and Deontic Logic. Oxford University Press.

[7] **[Lindner, Bentzen, and Nebel 2017]** Lindner, F.; Bentzen, M.; and Nebel, B. 2017.

The HERA approach to morally competent robots. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE/RSJ.

[8] [**Lindner and Bentzen 2018**] Lindner, F, and Bentzen, M. 2018. A Formalization of Kant's Second Formulation of the Categorical Imperative. arXiv:1801.03160v1

[9] **[Johnson R, 2014]** Kant's moral philosophy. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy.

[10] **[Kant I, (1785)]** Groundwork of the Metaphysics of Morals. Cambridge University Press.

[11] **[Manna and Nath 2021]** Manna, R, and Nath, R. 2021. Kantian Moral Agency and the Ethics of Artificial Intelligence. Vilnius University press. https://doi.org/10.15388/Problemos.100.11

[12] **[Morgenstern, Von Neumann, 1953]** Morgenstern O., Von Neumann J. 1953. Theory of games and economic behavior. Princeton University Press.

[13] **[Mougan et Brand 2024]** Mougan, C, and Brand, J. 2024. Kantian Deontology Meets AI Alignment: Towards Morally Grounded Fairness Metrics. arXiv:2311.05227v2.

[14] **[Nash, 1950]** Nash J. F. 1950. Equilibrium points in n-person games. Proceedings of the National Academy of Sciences of the United States of America, vol. 36, no 1, p. 48–49.

[15] **[Powers 2006]** Powers, T. M. 2006. Prospects for a kantian machine. IEEE Intelligent Systems 21(4):46–51.

[16] **[Singh, Lavanya 2022]** Singh, Lavanya. (2022). Automated Kantian Ethics: A Faithful Implementation. In: Bergmann, R., Malburg, L., Rodermund, S.C., Timm, I.J. (eds) KI 2022: Advances in Artificial Intelligence. KI 2022. Lecture Notes in Computer Science(), vol 13404. Springer. pp 187–208

[17] **[Vallée, Bonnet, de Swarte 2018]** Vallée,T; Bonnet, G, and de Swarte, T. 2018. Modélisation de valeurs humaines : le cas des vertus dans les jeux hédoniques. Revue d'intelligence artificielle – no 4/2018, 519-546

[18] **[Winfield, Blum, and Liu 2014]** Winfield, A. F.; Blum, C.; and Liu, W. 2014. Towards an ethical robot: internal mod[1]els, consequences and ethical action selection. In Mistry, M.; Leonardis, A.; M.Witkowski; and Melhuish, C., eds., Advances in Autonomous Robotics Systems. Springer. 85– 96