Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.38326

AI Value Judgment from the Perspective of Hume's Law: A Critical Inquiry

Jingxuan Luo

Department of Philosophy, Xiamen University, Xiamen, China
LuoJingxuan20039@outlook.com

Abstract. Value Alignment Engineering for AI aims to turn human normative demands into computable parameters. But there are problems with the conflict between ethical principles and disagreements in RLHF annotation. Therefore, this paper investigates the problem of whether AI can generate values with normative force. Hume's Law is a mistake in the reasoning about facts and values. A system to solve this problem is provided. The content of this paper is as follows: An AI system for answering questions and deriving values in accordance with a specific norm. This paper provides a formal reconstruction of Hume's Law as a constraint on AI value alignment; conducts architectural analysis to show how value judgments undergo 'factual degradation' in Transformer-based systems; and puts forward practical recommendations to transition from 'preference aggregation' to 'preference identification' in the RLHF pipeline. The first premise is to rebuild Hume's Law logically. It also responds to the earlier criticisms of "promises" by Searle and "thick ethical concepts" by Putnam to defend the authority of law. The second proposition is the statistics of Transformer-based AI. Therefore, even after introducing value judgments in the model (corpus input or RLHF fine-tuning), they are transformed into statistical facts of vectors and probabilities. Based on the reasons given above, this paper believes that there is a structural defect in the current value alignment. The limit is due to the value-gap problem in Hume's Law. Although AI has some real-world applications in ethical assistance, we should be careful not to have it independently generate values and assume moral responsibility.

Keywords: Hume's Law, Artificial Intelligence, Value Judgment, Value Alignment, Transformer Architecture

1. Introduction

1.1. Research background and problem statement

Currently, the Value Alignment Engineering of AI faces a number of practical problems. Reinforcement Learning from Human Feedback (RLHF) is a popular alignment method. Annotators are used to rank the model's outputs based on preference. However, the annotators were unable to agree on a preference, and the same behaviour was judged differently in different cultures and value systems. Despite annotation inconsistencies, AI has conflicts of ethical principles. Autonomous driving should weigh the protection of passengers and pedestrians. Medical AI needs to balance

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.38326

patient autonomy with the doctor's best interests. Therefore, the system has no reason to select one over the other. Therefore, the problem is whether AI can independently generate values of moral norms through data processing and optimisation.

Hume's Law is employed as the theoretical foundation in this paper to address the problem. (Throughout this paper, "Hume's Law" refers to the problem of moral reasoning that shows a logical gap between "ought" and "is", and it should not be confused with other problems raised by Hume). Therefore, this paper consistently applies "Hume's Law". Hume's Law originated from the idea of the British philosopher David Hume in A Treatise of Human Nature: one cannot logically derive "ought" (values) from "is" (facts) [1] . The philosopher Hare presented and named this idea "Hume's Law" [2]. According to Hume's Law, the division between facts and values will be adopted in this paper. Factual Judgments answer the question "what is", have truth values, and can be verified by observation. Value judgments are answers to the question of "what should be" and belong to evaluative judgments. They are conclusions of moral reasoning. The connection between the two is not based on reason. Therefore, Hume's Law offers a general framework to investigate the limits of AI's capacity for value judgments. If there is indeed a logical gap between facts and values, then any attempt by AI to reach normative conclusions based on the processing of only factual data must first address this gap.

1.2. Research question and main thesis

The two main reasons for this paper are:

P1 (Defense of the Validity of Hume's Law): Normative propositions cannot be deduced solely through logic from purely factual propositions. Firstly, this paper will introduce Hume's Law. It will answer the objections of Searle and Putnam. It will be shown that neither view violates the legality of the act.

P2 (Factual Constraints of AI Architecture): Value judgments in the model are all converted into some extent into factual data. Value judgments cannot retain their own status without factual judgments.

Based on the two above assumptions, this paper hopes to demonstrate that AI can answer questions and derive values in light of the given normative conditions. However, it cannot guarantee the logical reliability of the final value judgment (the validity of reasoning and their normative binding force) solely by means of data processing and optimisation. In addition, this paper will present the range of its ideas and will not ignore the practical advantages of AI for ethics.

1.3. Literature review and research gap

In contemporary AI ethics, Robert James M. Boyles conducted a number of studies. Systematically applied Hume's Law to various kinds of AMA (Autonomous Moral Agents) in his work, and has since provided a foundation for this field [3]. He found that in the event of a conflict among virtues, AI lacks a factual basis for determining which should be prioritised, and thus the system cannot operate effectively in such circumstances. Therefore, the application of Hume's Law to AI systems can be empirically verified.

Austin Spizzirri has also put forward the above problems in his work [4]. He indicated that the failure of the current main alignment method to meet Hume's Law is "structural, not an engineering defect", and also listed Hume's is-ought gap along with value pluralism and the extended frame problem as fundamental obstacles to alignment.

DOI: 10.54254/2753-7064/2026.38326

However, these studies have a common defect. Boyles directly applied Hume's Law to study AMA empirically. However, he did not provide many theoretical supports for the law. Spizzirri has studied the structural problems in alignment. However, he also offered little support for Hume's Law itself.

In light of the above, this paper adds to the discussion in the following two ways. First, it offers a logical reconstruction and systematic defence of Hume's Law. Second, it advances the study of AI architecture at the statistical level. It shows that the problem is due to the inherent mathematical characteristics of the architecture. Based on the above theoretical foundation, this paper will investigate the limitations of AI's value-judgment capability under the factual constraints imposed by AI architecture and the restrictions of Hume's Law.

1.4. Structure of the paper

The reasons for this are as follows. Section 2 presents a logical reconstruction of Hume's Law. It studies the extent of deductive and inductive reasoning in this problem. It also answers the traditional objections raised by Searle and Putnam. Therefore, the foundation P1 has been verified. Section 3 presents the statistical characteristics of the current mainstream AI system, Transformer-based large language models. Therefore, it believes that after the addition of values, they can be converted into some form of factual data by the model. This is Premise P2. Section 4 incorporates the two above. Therefore, the paper has the following conclusions. It also presents the theoretical and practical implications of value-alignment engineering.

2. Defending the validity of Hume's Law

2.1. Logical reconstruction of Hume's Law

To establish the reliability of premise P1, this paper will undertake a logical reconstruction of Hume's Law. It hopes to show that there are problems in both deduction and induction when getting from facts to values.

Hume's Law rules out the hypothesis of induction reasonably. Let Q be the factual proposition "The autonomous vehicle is about to hit a pedestrian ahead". Let R be the normative proposition "The autonomous system should brake immediately". R cannot be deduced from Q alone. If the system is to prioritise traffic efficiency, it cannot achieve R. For the purpose of derivation, we add a value V_1; that is to say, we add the premise of value: "the pedestrian's right to life takes precedence over any consideration of traffic efficiency", and then Q and V_1 together imply R.

Another way to address the deficiency is to add a particular case. For example, add F_0: "To hit the pedestrian ahead will cause injury.". R is still not recognised. Given the fact, "should immediately reduce speed" is a value judgment V_2: "given the foreseeable harmful consequences, the system has a practical duty to take feasible measures to prevent them". Only after adding V_2 can Q, F_0, and V_2 together infer R.

Therefore, to change it from a statement of fact to a statement of values, at least one value assumption needs to be added. Only the addition of factual judgments cannot resolve the logical gap in Hume's Law.

The path of induction is relatively less reliable than that of deduction in connecting facts and values. Induction is to draw a general rule from many particular cases by repetition. Its first defect is that it is probabilistic rather than necessary in logic. From the above, it can be seen that although "ten thousand people were injured after being hit by a car" is known, we cannot be certain that "any

DOI: 10.54254/2753-7064/2026.38326

person hit by a car will be injured". Secondly, induction occurs within the same logical category and derives universal propositions from a large number of singular ones. Only the quantifier changes in the inductive transition. No new concepts, relations or categories are added to those in the premises. Therefore, according to the factual judgments that do not contain "ought", induction cannot produce value judgments that contain "ought".

Therefore, it can be concluded that the logical limit set by Hume's Law cannot be exceeded through the accumulation of data. According to the system of deduction, to reach a conclusion of value judgment from the given facts, at least one such value judgment needs to be added to the premises.

2.2. Responses to classical objections

2.2.1. Response to Searle: institutional facts and the normative presupposition of promise

The main reasons for doubting the validity of Hume's Law have been proposed by Searle and Putnam. However, this paper does not believe that either is fundamentally invalid. Searle proposed the idea of an "institutional fact" and presented a five-step inference.

(1) Jones uttered the words "I promise to pay you five dollars."

(2) Jones promised to pay Smith five dollars.

(3) Jones placed himself under an obligation to pay Smith five dollars.

(4) Jones is under an obligation to pay Smith five dollars.

(5) Jones ought to pay Smith five dollars.

Searle thinks that it is a whole-logical and valid deduction [5]. The first defect of Searle's idea of an "institutional fact" is that "promise" already contains the implicit rule that "promises should be kept"; thus, Searle begins with a combination of norms and facts rather than pure facts [2]. Without the rule of thumb "promises should be kept", there would be no normative force for the institution of promises. This problem is further shown by the logical reconstruction of Hume's Law in this paper. Searle's inference finally leads to the value-judgment premise that "promises ought to be kept". His inference reaches a value judgment, and significantly introduces (or ends in the introduction of) the value judgment V through the inferential process. Therefore, according to Searle, this inference cannot resolve the logical gap shown by Hume's Law.

2.2.2. Response to Putnam: thick ethical concepts and the problem of overgeneralisation

According to Putnam, facts and values are in conflict. He proposes the idea of "thick ethical concepts" to provide support for the above argument [6]. Putnam's "cruelty" is a type of harm caused by one's behaviour that can be judged according to values and observed in practice. Therefore, he hopes to disprove Hume's Law or at least the purely logical form of it.

Putnam's observation has been verified in the first instance. In fact, there is a type of ethical idea that combines description and evaluation, and facts cannot be divorced from values. There are also some deficiencies in this observation: "thick ethical ideas" are only one kind of value, and not all of them. Putnam opted to introduce the idea of "thickness" and proposed the existence of another type of value judgment: "thin ethical concepts". These are typical examples of value judgments that were accepted prior to Putnam's proposal of the concept of "thickness". These ideas are not being presented; they are "good", "right" and "ought".

More importantly, the most effective scope of Hume's Law is precisely these "thin ethical concepts". When a value judgment is divorced from a specific situation, such as "theft is wrong", the

DOI: 10.54254/2753-7064/2026.38326

word "wrong" does not indicate any particular feature of the act of theft. It is only a negative normative designation. When the subject of the value judgment is abstract, such as "freedom is good", the word "good" does not refer to any observable attributes. Due to the above deficiencies, Putnam's critical strategy has lost its impact. There is no such content. Therefore, the logical gap cannot be solved by a fact-value entanglement strategy.

Therefore, Putnam's objection falls under the problem of overgeneralisation. He extends the attributes of some value judgments (thick ethical concepts) to all the value judgments in a group. Putnam's objection does not provide sufficient logical support for the norm according to the facts. Therefore, Hume's Law has been validated to some extent. Putnam's objection of "thick ethical concepts" does not invalidate the rule or the outcome of Hume's Law.

2.3. Summary of this section

In short, neither deduction nor induction can connect the problem of facts with that of values. The Value Proposition can be added at the end of the inference. Searle's "propositional" argument already presupposes a norm. Putnam's "thick ethical concepts" cannot cover all value judgments. Thus, P1 is introduced.

3. Factual constraints on AI architecture

3.1. Theoretical foundation: the difference between Strong Ethical AI and Weak Ethical AI

Next, this paper presents P2. To limit the scope of Hume's Law in the context of artificial intelligence, two interpretations of ethical AI proposed by Pertti Saariluoma [7] need to be distinguished. Strong Ethical AI can autonomously derive new normative conclusions from factual data. Weak Ethical AI refers to systems that "add values to technology". Humans provide the rules for values for the AI, and then it carries out ethical functions within that system.

Saariluoma believes that Strong Ethical AI can independently draw new moral norms based on data. According to Hume's Law, no algorithm can produce values or norms that are binding on behaviour by logic. Therefore, Strong Ethical AI is not practically feasible.

The case of Weak Ethical AI is more complex. Its reasons appear to be based on a combination of facts and values. It does not appear to be Hume's problem. But this is not actually the case. This paper will introduce the two ways in which values are added to AI.

3.2. Two pathways of value judgment input and their factual degradation

3.2.1. Input pathway one: corpus input and token vectorisation

Most modern large language models (LLMs) are built on the Transformer and related architectures. The basis of this architecture is a multi-head self-attention mechanism. At the core is the dot product and weighted sum of query vectors (Query), key vectors (Key), and value vectors (V_trans) [8].

That is to say, when providing the value judgments to the model in the form of a corpus, the processing method is as follows: map each token to a real-number vector in embedding space. Then compute the inner product of the vectors to obtain the statistical dependency at all positions in the sequence.

Therefore, in the application of "one ought not to steal", "ought" and "steal" can be considered examples. Both are subject to the rules of vector operations. The model does not have a function to recognise these as special linguistic expressions in the normative system. More precisely, the value

DOI: 10.54254/2753-7064/2026.38326

judgment "one ought not to steal" is formally converted into factual judgments about the probabilities of the tokens "not", "ought", and "steal" during processing, and then participates in subsequent data-wise statistical analysis.

3.2.2. Input pathway two: RLHF and other fine-tuning mechanisms

The second input path is for the fine-tuning mechanism of the model using value judgments, such as RLHF. Integration does not change the basic nature of fact processing. RLHF is, in fact, a problem of data analysis based on "noisy, subjective, and adaptively collected preference data" [9].

The annotator's normative judgment of evaluation is not added to the model as a category command. It is a scalar preference signal that has been affected by statistical noise. In other words, the norm is taken to be the people's desire.

3.3. The statistical nature of transformers and the argument for architectural universality

Research on Transformer mechanisms further supports its statistical nature. Actor and others have shown that the evolution of representations in Transformers can be viewed as a path to obtain optimal classification features; that is, the attention mechanism is essentially an implicit form of multinomial regression [10]. At present, academiaAt present, the academic community has reached a clear consensus regarding the nature of LLMs. Su gives a simple explanation: LLMs are "fundamentally statistical models because they are based on data and use stochastic generation processes". They are "a kind of autoregressive model that learns language by learning statistical patterns in human-generated text, rather than learning explicit linguistic rules or semantic foundations" [11]. Fokoue extends this view to all of artificial intelligence and believes that the theoretical and methodological foundation of artificial intelligence has always been, and continues to be, statistics [12].

Therefore, no matter how these inputs are provided, Transformer-based LLMs and other AI systems fail to preserve all the original values of the input. They are regarded as facts and should not be assumed to be different from the other factual data. Therefore, when value judgments are fed into AI and these value judgments are finally output, it seems that value judgments have been derived from a combination of value judgments and factual judgments in this process. In fact, the derivation may have employed a mixture of factual judgments and value judgments that no longer exist in their original form after being processed as factual judgments to some extent to produce the final value judgment. Thus, Hume's Law of the contrary may not apply,and no rational foundation is provided.

For example, other architectures besides Transformers may also be able to address the problem of Hume's Law. However, no matter how the architecture is modified,be it State Space Models (SSM), Mamba, or geometric algebra-based Versor,as long as the system performs input mapping (for tokens, vectors or other mathematical objects) to real-number vectors and then captures statistical relationships among the inputs through mathematical operations (dot products, weighted sums, matrix multiplications, etc.), the statistical nature of Transformers is not an arbitrary defect of that architecture. It is a general attribute of systems that process input data through algorithms.

3.4. Summary of this section

In short, whether they are added to the corpus or included in RLHF, they are converted into factual data for the AI, such as vectors and probabilities, and thus lose their independent normative force.

DOI: 10.54254/2753-7064/2026.38326

This is due to the statistical nature of algorithm processing and is architecturally universal. Thus, P2 is introduced.

4. Conclusion

4.1. Main conclusions

Therefore, the conclusion is obtained from the premises. P1 indicates that purely factual judgments cannot provide values through reason. P2 indicates that AI is carrying out some form of processing on the input value judgments to obtain factual judgments that cannot be guaranteed to differ significantly from other factual data. Therefore, Weak Ethical AI may produce, apply and even derive values based on the given normative system. It cannot ensure the logical correctness of the final value judgment by itself through data processing and optimisation.

4.2. Theoretical and practical significance

The primary contribution of this paper is to offer a logical reconstruction and continuous defence of Hume's Law. It can be seen that, at present, "value alignment" engineering faces not only issues in the execution of engineering but also structural constraints. The pursuit of machine values should not be based on increasing the quantity of data and models. Acknowledge the actual advantages of artificial intelligence and address the deficiencies. We should be careful about the capacity of autonomous value judgment by AI and whether it can be independently held morally and institutionally responsible.

Based on the above analysis, the annotation discrepancies in RLHF should not be treated as statistical noise. They should be considered empirical expressions of the difference in normative foundations. Engineering practice can shift from "preference aggregation" to "preference identification". Record the types of value presuppositions in the annotations for disagreements. Do not delete the normative systems during model training; instead, keep several. Display the system's output and several conclusions at the same time, and then have a person decide. Therefore, the function of AI will be to generate "normative output", and human decision-makers will remain responsible for the final decision.

4.3. Research limitations and future directions

The two reasons are as follows. First, it points out that AI has been applied in the field of ethics, but it has not explored the limits of such application. Second, it mainly studies Transformer-based LLMs. Discussions of non-Transformer architectures (SSMs, Mamba, etc.) are currently limited to logical deduction and have not yet been empirically verified.

Future research can proceed in two directions. First, only one AI system is included in the current analysis. Further research can explore the application of Hume's Law in multi-agent systems. When several AI systems are in communication and have different values, can a new normative consensus be formed among them through the exchange of factual information? Second, research can explore explicit methods for the normative basis of human value input in "value alignment" engineering. It can identify the constitutive rules in the same way that "promises ought to be kept" relates to "promise". Moreover, new architectures with fundamentally different underlying logic may appear in the future, and the applicability of the present discussion will be correspondingly limited.

DOI: 10.54254/2753-7064/2026.38326

References

[1] Hume, D. (1980). *A treatise of human nature (Vol. 2)* (W. Guan, Trans.). Commercial Press. (Original work published 1740)

[2] Hare, R. M. (1964). The promising game. *Revue Internationale de Philosophie, 18*(70), 398–412.

[3] Boyles, R. J. M. (2025). Conjectures on the possibility of recasting Hume's guillotine for hybrid-designed artificial moral agents. *Philosophia: International Journal of Philosophy, 26*(2), 195–212. https://doi.org/10.46992/pijp.26.2.a.3

[4] Spizzirri, A. (2025, revised 2026). The specification trap: Why static value alignment alone is insufficient for robust alignment (arXiv:2512.03048v5). *arXiv*. https://doi.org/10.48550/arXiv.2512.03048

[5] Searle, J. R. (1964). How to derive "ought" from "is". *The Philosophical Review, 73*(1), 43–58. https://doi.org/10.2307/2183201

[6] Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Harvard University Press. https://doi.org/10.2307/j.ctv1pdrpz4

[7] Saariluoma, P. (2020). Hume's guillotine in designing ethically intelligent technologies. In T. Ahram, W. Karwowski, A. Vergnano, F. Leali, & R. Taiar (Eds.), *Intelligent Human Systems Integration 2020: Proceedings of the 3rd International Conference on Intelligent Human Systems Integration (IHSI 2020)* (pp. 594–599). Springer. https://doi.org/10.1007/978-3-030-39512-4_92

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3295222.3295349

[9] Liu, P., Shi, C., & Sun, W. W. (2026). Reinforcement learning from human feedback: A statistical perspective. arXiv. https://arxiv.org/abs/2604.02507

[10] Actor, J. A., Gruber, A., & Cyr, E. C. (2025). Deriving transformer architectures as implicit multinomial regression (arXiv:2509.04653v2). arXiv. https://doi.org/10.48550/arXiv.2509.04653

[11] Su, W. J. (2026). Do large language models (really) need statistical foundations? *The Annals of Applied Statistics, 20*(1), 724–743. https://doi.org/10.1214/26-aoas2151

[12] Fokoué, E. (2025). No intelligence without statistics: The invisible backbone of artificial intelligence (arXiv:2510.19212v1). arXiv. https://doi.org/10.48550/arXiv.2510.19212