AI and Ethics (2022) 2:449–461
https://doi.org/10.1007/s43681-021-00091-y

ORIGINAL RESEARCH

Coarse ethics: how to ethically assess explainable artificial intelligence

Takashi Izumo[1] · **Yueh-Hsuan Weng**[2]

Received: 26 June 2021 / Accepted: 30 August 2021 / Published online: 12 September 2021

Abstract
The integration of artificial intelligence (AI) into human society mandates that their decision-making process is explicable to users, as exemplified in Asimov's Three Laws of Robotics. Such human interpretability calls for explainable AI (XAI), of which this paper cites various models. However, the transaction between computable accuracy and human interpretability can be a trade-off, requiring answers to questions about the negotiable conditions and the degrees of AI prediction accuracy that may be sacrificed to enable user-interpretability. The extant research has focussed on technical issues, but it is also desirable to apply a branch of ethics to deal with the trade-off problem. This scholarly domain is labelled *coarse ethics* in this study, which discusses two issues vis-à-vis AI prediction as a type of evaluation. First, which formal conditions would allow trade-offs? The study posits two minimal requisites: adequately high coverage and order-preservation. The second issue concerns conditions that could justify the trade-off between computable accuracy and human interpretability, to which the study suggests two justification methods: impracticability and adjustment of perspective from machine-computable to human-interpretable. This study contributes by connecting ethics to autonomous systems for future regulation by formally assessing the adequacy of AI rationales.

Keywords Explainable AI · AI ethics · Robot ethics · Human–robot interaction · Autonomous vehicles

1 Introduction

Artificial Intelligence (AI) is increasingly used in discrete social domains; hence, it is becoming progressively crucial to refer to its predictions [10, 37]. Human beings should be able to apprehend AI decision-making because interpretability contributes to the detection and rectification of partiality or bias [2, 53]. The term *explainable AI* (XAI) was posited in 2016 by the Defense Advanced Research Projects Agency[1] (DARPA) [18] to improve interpretability. Thereafter, diverse models were proposed as surveyed by Arrieta et al. [2], Guidotti et al. [17], Hall [20] and Molnar [36]. Such models include SHapley Additive exPlanations [33, 36] or Local Interpretable Model-agnostic Explanation (LIME) [36, 40]. This issue applies to human–robot interactions (HRI) too, as it has been consistently queried whether robots could share ethics or at least display ethical similarities with human beings since the annual IEEE RO-MAN series in 1992 [13] and whether their ethics could be easily comprehensible in the manner of Asimov's Three Laws of Robotics [52]. In addition, the study of the Law of Algorithms would also incorporate issues contemplating ethically acceptable standards to balance AI accuracy against interpretability [5].

However, previous studies have demonstrated that the transactions between the accuracy of AI predictions and human interpretability represent a trade-off [2, 10, 19, 36, 47]. In other words, accuracy could be relinquished in the quest for easier human understanding. This trade-off problem is an issue of ethics too [10, 23]. Therefore, the conditions and degrees to which the accuracy of a prediction may be sacrificed in favour of human interpretability must be established. In other words, the instances in which it would be permissible for XAI to roughen its own predictions must be elucidated.

Historically, this ethical question was deemed so important that Thomas Hobbes [24] postulated a similar problem in his *Leviathan*, concluding that his natural law theory

✉ Takashi Izumo
izumo.takashi@nihon-u.ac.jp

[1] College of Law Graduate School of Law, Nihon University, Tokyo, Japan

[2] Frontier Research Institute for Interdisciplinary Sciences, Tohoku University, Sendai, Japan

[1] The announcement is identified with DARPA-BAA-16–53.

was difficult but became interpretable for everyone using the so-called Golden Rule: 'Do not that to another, which thou wouldst not have done to yourself'. The present study's contemplation of the traditional concerns about the trade-off between accuracy and interpretability would contribute to ethics as well as to the future regulation of XAI. The research domain delving into this issue is labelled *coarse ethics* (CE)[2] in this study. This branch of inquiry encompasses AI ethics and human–robot interactions (HRI) [4, 6, 13, 44].

Section 2 addresses previous research and recent gaps in this issue, along with the methodology, assumptions and limitations of this study. Section 3 defines CE in comparison to traditional ethics—especially to the Kantian one. Section 4 introduces two requisites of adequately high coverage and order-preservation to elucidate the concept of normative coarsening, starting with the question of how humans perform such simplification in everyday life, followed by rigorous definitions and formulations. XAI's relevance to this issue is shown in several topics. In addition, Sect. 4 discusses how to justify trade-offs and posits two justification methods: justification for impracticability and perspective adjustment. After all the definitions and formulations, Sect. 5 applies CE to the case of self-driving car. Section 6 concludes this study.

2 Previous research and methodology

2.1 Previous research

In the tradition since Aristotle, the ethical concept of responsibility has required explainability about the decision-making process [10, 35]. Humans should explain their decision-making process, and to be more precise, they should find a good explanation to justify their behaviour [10]. If this rule applies not only to human beings but also to AI, then AI also needs to select a good explanation for its own prediction [2, 8, 10, 19, 35, 47]; this means not only disclosing the software code but also explaining its decision to stakeholders [10, 47]. Although each explanation would depend on its social context and AI could be discussed in a different context than humans [8, 35], the purposes of XAI are similar to the ones of human explanation, that is, to make its behaviour intelligible [19], make its predictions trustworthy [10] (as the High-Level Expert Group on AI of European Commission presented the Assessment List for Trustworthy Artificial Intelligence in 2020 [47]) and avoid bias [2, 53]. The facets to which AI should be held accountable must be determined during the development phase, and this task relates to the domain of Explainable AI Planning (XAIP) [15].

It is widely accepted that there is a trade-off between accuracy and interpretability in AI decision-making [2, 10, 19, 36, 47]. This trade-off is manifested in the tension that the performance of AI must be reduced to make it intelligible and trustworthy or avoid bias, and the two parameters of accuracy and interpretability cannot be maximised concurrently [2, 19]. Furthermore, there is no such uniform criterion by which accuracy is always more important than interpretability or vice versa [10]. However, some agencies, such as DARPA [18] and the Joint Research Centre of EU Commission [21], are forbidding AI without any accountability at all; thus, the trade-off problem is not only a technical but also an ethical issue.

Explainability is normally ensured when AI adopts a transparent model-like decision tree, but such a model does not perform as well as deep learning, which uses neural network [2, 10, 19, 35, 36]. Therefore, a procedure called 'open the black box' is being explored to first make predictions with deep learning and then explain them with a transparent model [10]. This is the fundamental conception of XAI, and various technical methods have now been established for this procedure [2, 35, 36]. As Arrieta et al. [2] summarised, the discussion on XAI can be categorised into two stages: model design and post hoc techniques.

First, an important aspect of model design involves the selection of one from several transparent machine-learning models. Arrieta et al. [2] state that to be interpretable, all transparent models should conform to one or more of the three conditions of simulatability, decomposability and algorithmic transparency. Arrieta et al. [2] list six models that satisfy the requisite stipulations: linear/logistic regression, decision trees, *k*-nearest neighbours, rule-based learning, general additive models and Bayesian models. Arrieta et al. [2] further demonstrate how the trade-off between prediction accuracy and human interpretability occurs in different models.

Second, post hoc techniques are required when non-transparent black box models are chosen, especially when Deep Neural Network (DNN) is adopted [2]. There are some famous techniques, for example, LIME, which attempts to locally approximate a black box model into a linear regression model [36]. Inductive Logic Programming (ILP) can be also applied to explain the original black box model [38].

[2] The term *coarse ethics* is coined by the authors, but its roots can be traced to the theological probabilism of the early-modern Catholic church: in selecting one of two opinions, priests were allowed to follow the less reasonable one, leading to moral tolerance. For more on this theory, see Schwartz D (2014) Probabilism Reconsidered: Deference to Experts, Types of Uncertainty, and Medicines. *Journal of the History of Ideas* 75(3), pp. 373–393 [43]. Further, the adjective *coarse* is inspired from coarse geometry, a branch of pure math researching ways in which the structure of space can be described from a distance.

In addition, as etiquette and conversational wit demand specific applications such as Siri to converse in polite manner, communication is an issue for XAI and HRI as well. The diversity of explanations can be exemplified by the life story book, which Setchi et al. [44] suggest for reminiscence therapy by robots. Another example is verbal guidance for a sit-to-stand support system [46].

Since various ideas have been proposed, it is believed that currently, there is no common means to evaluate and measure XAI [19]. Therefore, this study starts with the following question:

Are there any common requirements that all types of XAI must meet? If they exist, where is the lower bound of trade-offs and how can it be assessed?

There are four current solutions to this question. First, the concept of human rights plays a major role, and the so-called 'human-centred approach' [9] or 'human rights-centred approach' [55] possibly forms a minimum requirement. For example, if a model design poses a 'high' or 'very high' risk to human rights, it could be possible to impose a legal obligation on the developer to reduce the risk [28, 39, 55]. If human rights form the lower bound of trade-offs, it is not permissible to improve performance at the expense of human rights, and conversely, it is justifiable to worsen performance on the grounds of guaranteeing human rights.

By contrast, the rest of the solutions do not match the human rights-centred approach. One of them is posthumanism, which argues that not only human beings but also other animals and abiotic actors should be stakeholders [22]. If the second solution is correct, then the lower bound of trade-offs should be higher than in the human-centred approach because animal rights and abiotic actors' rights would also be considered, and this consideration would require people be more ethical against the environment [22].

The third solution is pragmatism, which argues that trade-offs can be dealt with only retrospectively; alternatively, trade-offs should not be interpreted as including rational, cognitive or utilitarian valuations, but they should be regarded as justification after the fact [23]. In this case, trying to set a lower bound on trade-offs in advance would be pointless.

Lastly, Rudin [41] has asserted that trying to explain a black box is not a good option for high-stakes decisions, and a transparent model should be adopted from the outset. As Rudin [41] has denied that the trade-off problem exists, the aforementioned question would stand invalid.

This study does not follow the third and fourth solution, that is, it supposes that opening the black box is technically and ethically possible, and these trade-offs should be assessed in advance—at least for human rights issues.

2.2 Recent challenges and gaps

The first and second solutions provide only a rough guide to the minimum requirements that XAI should satisfy; therefore, many details are yet to be defined, such as the educational level of users that XAI should target [8], possibility of training XAI in a certain direction [10] and true meaning of accountability [30]. Nevertheless, one of the most important challenges is formalisation [3, 8, 31]. Even if human rights may form the lower bound of trade-offs, the question of whether a certain trade-off is congenial to human rights or not can be answered merely on a case-by-case basis as the resolution of conflict between rights and collective interests depends on the affected stakeholders [55]. Of course, the formalisation of ethics is thriving, and deontic logic is one of the most popular examples [27]. Limited to XAI, there are studies on the formalisation of explanation itself as well. For example, Chrisley [9] has applied deontic logic directly to XAI. General Ethical Dilemma Analyzer (GenEth) [1] and Value-driven Agent (VDA) [32] have been developed for ethically justifying AI predictions with ILP. Furthermore, directed acyclic graphs can be applied to analyse algorithmic fairness [3].

However, these models do not address how to formalise trade-offs, or in other words, how to formally simplify predictions by black box models; for example, they use only integer values [1, 32] or abstract concepts such as 'gender' [3] for reasoning. If the complexity of the real world is considered, the ethical value of each action should be more diverse and have finer differences, and it is not plausible to assume that DNN would output only integer values or classify people into a few gender groups.

As for formalising trade-offs itself, there is an attempt called Key Ethics Indicators (KEIs) [31], which proposes a series of steps to compare the decision maker's values with the calculated results of each trade-off method and to select the model in which the latter best reflects the former. Lee et al. [31] assume, however, that fairness metrics are subjective and that their measurement varies from one case to another, in other words, the purpose of KEIs is not to establish general forms of trade-offs but to create a customised measurement in each model.

Therefore, the fundamental question of under what form ethical values can be simply qualified or quantified through trade-offs (e.g. under what form these values can be quantified only in terms of integer values) is left unanswered. This study attempts to fill this gap.

2.3 Methodology, assumptions and limitations

To fill this gap, this study makes the following three assumptions:

- Human ethical judgement cannot perfectly grasp actual values and contain qualitative or quantitative simplification. This simplification itself can be an issue of ethics, and it also occurs in opening the black box, that is, output data by a black box model cannot be perfectly reproduced in a transparent model.
- Humans can distinguish between reasonable and unreasonable simplifications. There are such things as formal criteria (i.e. formality that distinguishes between them); this formality can also be applied to the trade-off problem of XAI.
- A field of ethics dedicated to the clarification of the formality of trade-offs can be established. This field does not aim to identify a strict ethical solution, but rather to clarify under what formality a strict ethical solution, e.g. a prediction by DNN, can be roughly treated by a transparent model.

Based on these assumptions, this study provides a foundation for the trade-off problem through a partial critique of Kant's ethics. It focuses on Kant because his ethics and its usefulness are often mentioned and approved in the context of AI ethics [9, 10, 37]. Using this critique, this study finds two formal requisites for XAI, namely, adequately high coverage and order-preservation; however, because the theme of the study is the formality of trade-offs, the fulfilment of this formality does not dictate which trade-off is the best. Therefore, instead of discovering the best trade-off, this study introduces some inappropriate ones, providing a list of failures in trading-off.

There are two limitations to this study. First, it does not guarantee that all formal minimum requirements will be enumerated. This study identifies only the aforementioned two requirements that should be met by XAI. Second, the formality of simplification within trade-offs may not directly answer the question of what trade-off should be practically chosen. However, the task of opening the black box of AI for citizens to review AI must have a certain formality, or else the method of opining the black box itself could be arbitrary. For the claim that developers sacrificed accuracy of AI prediction for its transparency to be valid, this trade-off needs to satisfy a predetermined formality, even if the mere fulfilment of this formality may not be enough for this sacrifice to be approved in the end.

3 Coarse ethics

3.1 The definition of coarse ethics

A concept must be defined before it is explored. CE is reflectively defined in this study by excluding the following premise that seems to be widely adopted in traditional ethics [14]: it is required to assume that ideals exist, and if one of two choices approximates an ideal more than the other, then the former should be evaluated higher than the latter. It is easy to illustrate this assumption citing the case that a person who lies only once per year should be evaluated higher than another who lies every day if the maliciousness of their lies is equal, because the former is closer to the ideal of honesty.

This premise can be found, for example, in Kantian ethics [16]. He explains his own methodology as follows:

> Virtue is always *in progress* and yet always starts *from the beginning*.–It is always in progress because, considered *objectively*, it is an ideal and unattainable, while yet constant approximation to it is a duty. That it always starts from the beginning has a *subjective* basis in human nature, which is affected by inclinations because of which virtue can never settle down in peace and quiet with its maxims adopted once and for all but, if it is not rising, is unavoidably sinking. [29]

It is assumed in this text that an ideal represents the extreme of a smooth normative evaluation function and that humans are capable of asymptotically approaching it; rather, they are obligated to asymptotically approach it. Kant does not offer a specific example, but the above example of two liars is useful here: each person should approach the ideal of honesty even though the individual cannot become completely honest; hence, the person who lies only once every year should be appraised higher than the other who lies every day, because it is not acceptable to refer to the proverb that a miss is as good as a mile.

This asymptotic approach, however, runs into a problem; a smooth evaluation cannot be directly applied. In considering the issue of global warming, for example, it is easily understood that eating vegetables contributes more to the reduction of greenhouse gases than eating meat [42]. However, it is difficult to ascertain the type of vegetable combination that will reduce greenhouse gases the most, and ethical judgements will become impossible if people are instructed to weigh the daily increases and decreases of greenhouse gases to the nanogram every day. In short, while Kant states that people should endeavour to raise their ethical positions to stop falling, we cannot accurately estimate whether we are slightly rising or falling at any given moment.

Kant notes this difficulty and offers a short reflection on such matters:

> But that human being can be called fantastically virtuous who allows *nothing to be morally indifferent* (*adiaphora*) and strews all his steps with duties, as with mantraps; it is not indifferent to him whether I eat meat or fish, drink beer or wine, supposing that both agree with me. Fantastic virtue is a concern with petty details which, were it admitted into the doctrine of virtue, would turn the government of virtue into tyranny. [29]

Table 1 The first evaluation method

Score	0	1	2	3	…	97	98	99	100
Evaluation	0	1	2	3	…	97	98	99	100

Table 2 The second evaluation method

Score	0–59	60–69	70–79	80–89	90–100
Evaluation	Inadequate	Fair	Good	Very good	Excellent

This description can be grounded in the daily experience that it is impossible to make uninterrupted ethical judgements about every trivial action. However, Kant does not mention here how a phenomenon can be identified as indifferent. Further, the example Kant cites of meat and fish is inappropriate for modern society because given global warming and veganism, the choice of meat eating cannot be said to be ethically indifferent. Such discrepancies can occur between AI and its users; that is, a choice could appear indifferent for users but different for AI. In such an instance, the trade-off would be problematic if AI explains something different as indifferent by sacrificing accuracy for human interpretability, because a potential risk embedded in the difference is kept hidden from the users. In other words, the concept of 'indifferent' (in Greek: *adiaphora*) is not self-evident; rather, this notion depends on what should be disregarded as indifferent.

CE can contribute to the resolution of the two problems of the asymptotic approach: the impossibility of smooth evaluation and the relativity of indifference by relinquishing the asymptotic approach and by permitting regulators, judges and agents to rough up their regulations, judgements or actions. This paper henceforth uses the term 'evaluation' for the abovementioned normative activities for the sake of simplicity. CE permits the evaluation of two objects as indifferent even if one of the items is actually or possibly closer to the ideal. The conception of CE is potentially related to every topic concerned with the limits of human understanding, for example, how to teach children or how to explain technology to non-experts. As the sections that follow evince, such a coarse approach facilitates answers to questions about the formal conditions within which XAI is allowed to sacrifice its own accuracy for human interpretability and to what degree.

3.2 Coarsening and refinement

3.2.1 Illustrative overview

There are discrete ways to perform rough evaluations. This section offers a simple example and introduces three special symbols. Teachers are obligated to assess students according to their scores. How the evaluation is achieved depends on decisions made by individual teachers as follows:

3.2.1.1 The first evaluation method (E_1) Students take a 100-point test in one-point increments (i.e. without decimal) and are evaluated in the order of their scores Table 1.

3.2.1.2 The second evaluation method (E_2) Students take a 100-point test in one-point increments and are evaluated as 'inadequate' if they score between 0 and 59, 'fair' if they secure tallies between 60 and 69, 'good' if their marks are computed between 70 and 79, 'very good' if they accrue points between 80 and 89 and 'excellent' if their scores are calculated between 90 and 100 Table 2.

3.2.1.3 The third evaluation method (E_3) Students take a 100-point test in one-point increments, and those who attain 60 or more points pass the test, whereas those aggregating points under 60 must retake the test as a penalty Table 3.

In such a case, E_2 is a rough version of E_1, and E_3 is a rough version of E_2. For now, the following simple arrangement is introduced: if two students s_i and s_j, who are evaluated differently in more fine evaluation (i.e. $s_i > s_j$ or $s_i < s_j$ according to their scores), are evaluated identically in coarse evaluation, then s_i is said to be coarsely equivalent or coarsely equated to s_j, and this relation is denoted as $s_i \approx s_j$ in the coarse evaluation. In CE, the symbol $\approx$ denotes normative coarse equation, and the following formulation is introduced:

$$o_i \approx^{\substack{\text{Evaluation\_method} \\ \text{Value}}} o_j$$

For example, if a student (s_1) scores 81 and another student (s_2) scores 82, they are distinguished if E_1 is applied but receive the same rating of 'very good' if E_2 is employed. Therefore, in terms of E_1, they should be assessed as $s_1 < s_2$, are deemed coarsely equivalent according to E_2 conditions and denoted as follows:

$$s_1 \approx^{\substack{E2 \\ \text{VeryGood}}} s_2$$

Further, if a student (s_3) scores 79 and another student (s_4) obtains 80, they are distinguished when both E_1 and E_2 are applied: s_3 gets 'good' and s_4 'very good' according

Table 3 The third evaluation method

Score	0–59	60–100
Evaluation	Failure	Pass

to E_2; however, both are simply adjudicated as passing the examination if E_3 is employed and may be represented as

$$s_3 \approx^{E3}_{\text{Pass}} s_4$$

In addition, if the relation between o_i and o_j is unequal, a similar formulation may be applied as follows:

$$o_i <^{\text{Evaluation\_method}}_{\text{Comparative\_value}} o_j$$

For example, s_1 and s_2 can be described as follows:

$$s_1 <^{E1}_{\text{Higher}} s_2$$

3.2.2 Coarsening and refinement

In the three evaluation methods described above, E_1, E_2 and E_3 are partially compatible with each other. For example, s_1 can translate marks obtained in an examination from E_1 (81) to E_2 (very good) to E_3 (pass). When an evaluation method E_j is a rough version of another, such as E_i, E_i is coarsened into E_j and is denoted as $E_i \rightsquigarrow E_j$ (coarsening). Thus, the relationship between E_1, E_2 and E_3 can be represented as $E_1 \rightsquigarrow E_2 \rightsquigarrow E_3$, and the notation $E_1 \rightsquigarrow E_3$ is possible as well. Conversely, if an evaluation method E_i is a more honed version of another, such as E_j, E_j is refined into E_i and is denoted as $E_j \curvearrowright E_i$ (refinement). For example, the above four students may be described as follows:

$$s_3(79) <^{E1}_{\text{Higher}} s_4(80) <^{E1}_{\text{Higher}} s_1(81) <^{E1}_{\text{Higher}} s_2(82)$$

$$\rightsquigarrow s_3 <^{E2}_{\text{Higher}} s_4 \approx^{E2}_{\text{VeryGood}} s_1 \approx^{E2}_{\text{VeryGood}} s_2$$

$$\rightsquigarrow s_3 \approx^{E3}_{\text{Pass}} s_4 \approx^{E3}_{\text{Pass}} s_1 \approx^{E3}_{\text{Pass}} s_2$$

In this case, coarsening is possible, but refinement is impossible. The best student (s_2), who scores 82 points, understands that the evaluation is 'very good' on E_2 and 'pass' on E_3 (i.e. the possibility of $E_1 \rightsquigarrow E_2$ and $E_1 \rightsquigarrow E_3$); however, a student only aware of having passed an exam cannot apprehend without additional information how many points were earned (the impossibility of $E_3 \curvearrowright E_1$). The possibility of refinement is not always guaranteed; thus, the initial assessment should be carefully selected. This study focuses on coarsening because of this unidirectional tendency.

3.2.3 Two requirements for adequate coarsening

The illustration offered thus far must be more rigorously explained to serve the purpose of academic contention. This study does not intend to argue that a null score is coarsely equated with a full score for lazy students. In other words, the concept of coarsening does not imply the disarrangement of an evaluation. Therefore, the question of under what formality the changing of an evaluation denotes adequate coarsening should be answered. This study posits two formal conditions for adequate coarsening: adequately high coverage and order-preservation.

3.2.3.1 The first requirement: adequately high coverage

E_j is said to cover E_i if and only if a rough evaluation method such as E_j encompasses one or more objects evaluated by the original E_i. The degree to which E_j encompasses E_i is called *coverage*. The coverage from E_j to E_i is said to be *full* if and only if E_j incorporates all the objects of E_i and does not include extra objects. Further, the coverage extended from E_j to E_i is labelled *over-coverage* if and only if E_j includes objects not evaluated by E_i. For adequate coarsening, the coverage offered by the rough evaluation method E_j to its original E_i is full or at least sufficiently high from some perspective. Adequately high coverage may include over-coverage in some instances.

Full coverage is desirable for adequate coarsening, which guarantees that every object evaluated by E_i can be found in E_j and vice versa. Such full coverage would free a judge from the trouble of not finding a rule for some objects. However, this study does not consider full coverage a necessary condition for adequate coarsening because of the costs of its satisfaction. Rather, coarsening is regarded as adequate when the coverage from a rough evaluation method to its original technique is sufficiently high from some standpoint. Thus, full coverage is characterised simply as the most evident case of adequately high coverage, and for the same reason, over-coverage (i.e. over-regulation) is also acceptable.

This mitigation helps several XAI models, such as Ribeiro et al.'s [40] LIME. LIME primarily trains local surrogate models to explain individual predictions and approximates the predictions through an interpretable surrogate model (e.g. by linear regression [35]). For instance, after a deep-learning model predicts the existence of a cat in an image, it can explain the reason for the prediction by focussing on contributory features such as long whiskers, triangular ears and distinctively tilted eyes, even though the original black box appears to use the whole image. The concept of adequately high coverage can make the local approximation ethically justified, and at the same time, it can help its users avoid deception by low coverage, as in the instance when a person wearing a cat mask is recognised as a cat.

Tjoa et al. [49] surveyed examples of LIME application to medical diagnosis. Once such identification concerned the prediction of influenza by highlighting the importance of certain symptoms (headache, sneeze, weight and no fatigue). This example can be formulated in CE as follows:

$$\text{weight} < \overset{\text{Flu\_diagnosis}}{\underset{\text{More\_contributes}}{}} \text{no fatigue} < \overset{F.d.}{\underset{M.c.}{}} 0 < \overset{F.d.}{\underset{M.c.}{}} \text{headache} < \overset{F.d.}{\underset{M.c.}{}} \text{sneeze}$$

This inequation demonstrates the degree to which each symptom contributes to the prediction of influenza. In this example, sneezing and headaches contribute positively, but no fatigue and weight contribute negatively to the likelihood. Does the diagnosis adequately encompass the original black box? It is difficult to establish a unique criterion for this estimation; thus, the present study will probe the question of who should bear the burden of proof of adequately high coverage. For example, if a transparent model with LIME does not mention a symptom to which not only a deep-learning model but also human doctors usually pay attention, then it should additionally explain why the important feature was ignored after the local approximation. The key point of this analysis is that such coarsening requires stronger justification than full coverage, and the burden of justification should be imposed on AI or its developer, not on users. Conversely, the diagnostic AI has satisfied its own black box accountability if human doctors also use the same four signs to diagnose influenza.

3.2.3.2 The second requirement: order-preservation Assuming the order between an object o_i and another object o_j is $o_i = o_j$ according to the original evaluation, this order may be termed *properly preserved* if and only if the order $o_i = o_j$ is maintained by a rough evaluation. When the order of the two objects is $o_i < o_j$ according to the original evaluation, this order may be called properly preserved if and only if the order is sustained as $o_i < o_j$ by the rough evaluation. However, this order is called *coarsely preserved* if and only if the order is changed into $o_i \approx o_j$. Adequate coarsening mandates that the order of all objects is properly or at least coarsely preserved; the order must not be reversed for any object.

The following example may be contemplated vis-à-vis the order-preserving stipulation for XAI. A fintech company financed B, but not A. Hence, A queried the rationale for the rejection from the financing AI of this company. The AI selected only one reason for the rejection and explained it as 'Sorry, you have too little security to allow lending'. However, this explanation fails to preserve the order of clients if B has no security and was able to borrow money merely by evidencing talent. The original evaluation was that B had more investment suitability than A when all things considered ($A < \overset{\text{Total\_value}}{\underset{\text{More\_suitable}}{}} B$), but the rough explanation meant $A > \overset{\text{Security}}{\underset{\text{Less\_suitable}}{}} B$. This reversal makes A distrust the AI because the rough clarification that was tendered does not match the order between A and B for financing. The AI must thus enumerate all reasons until A is persuaded of the order, $A < B$.

3.2.4 Definition: adequate coarsening

The exact definition of the special symbol $\rightsquigarrow$ can now be tendered, as noted below. The original evaluation E_i is said to be adequately coarsened into E_j if and only if both the requirements of adequately high coverage and of order-preservation are satisfied; only this adequate relation is represented as $E_i \rightsquigarrow E_j$. This definition enables the avoidance of dishonest simplifications and relieves CE from arbitrary revaluations.

4 Two arguments for coarse justification

4.1 Outline

The introduction of the two fundamental requirements of adequately high coverage and order-preservation is merely the starting point of CE. There exist infinite patterns of evaluation that satisfy both the requirements, as in the case of the mathematics examination. Therefore, each evaluator calls for additional arguments to decide which evaluation method is relatively better than the others even if the best one is not easy to find. In this study, the process to validate the selection is called *coarse justification*, and two possible arguments are proposed: impracticability and perspective adjustment.

First, the argument of impracticability implies that human users cannot always follow AI predictions because of their understanding or response speeds; AI can reference this fact to justify its own rough explanations. Second, the contention of perspective change is concerned with the viewpoint of evaluators. A student who earns 90 points in a mathematics examination should be assessed higher than another who attains 70 points; however, this difference would become trivial if they met at a class reunion after 50 years. The trivialisation occurs as the perspective alters from a student's viewpoint to a retiree's one. It also occurs during the perspective switch from AI to user or user to AI. There could also be an instance in which overfitting the AI to the training data sets makes the AI think that the mathematics examination results of 50 years back remain somehow relevant, but this particularised evaluation should be restrained from the standpoint of human beings. These arguments are analysed in the two sections that follow.

4.2 Coarse justification for impracticability

4.2.1 Basic idea

The first justification method for coarsening an evaluation concerns human impracticability or the experience that human performance is practically limited to some degree. For example, a navigating AI for a car driver says, 'Please slow down to 87.325 km/h!' This act may not be practicable for the driver even if the instruction is correct according to the environment sensed by the navigating AI. The AI should thus issue an instruction such as 'Please slow down to around 85 km/h!' or switch to autonomous driving mode. According to CE, the instruction instance can be formulated as follows:

$$85\ \text{km/h} < \overset{\text{AI–based evaluation}}{\text{Safer}}\ 87.325\ \text{km/h}$$

$$\rightsquigarrow 85\ \text{km/h} \approx \overset{\text{Human–based evaluation}}{\text{Safe}}\ 87.325\ \text{km/h}$$

Such coarsening can be justified both technically and ethically through the famous Latin proverb: 'Nobody is obligated to do the impossible' (Latin: *Ad impossibilia nemo tenetur*).[3] It should be noted, however, that CE does not allow for all kinds of coarsening for impracticability. If the protection of human life forms the lower limit of trade-offs, rough explanations that could lead to personal injury should be prohibited. In other words, XAI faces the problem of trade-off between safe human–robot interaction and effective task execution, which Sidobre et al. [45] mentioned in the context of HRI. In this case, as Contissa et al. [11] discussed, the trolley problem in automated driving must be examined additionally.

4.2.2 Definition

E_{i} may be coarsened to E_{j}, regarding $d_{\min}$ as its lower limit, if and only if the minimal difference that an evaluator can control is $d_{\min}$, but the difference that an original evaluation method E_{i} adopts is smaller than $d_{\min}$. Such coarsening can be validated because the original norm is not practicable and is therefore termed *coarse justification for impracticability*.

The discussion so far can be related to informed consent in medical care. For example, the ideal concept of informed consent mandates a doctor to accurately explain the treatment to the patient; however, this ideal concept must be coarsened for children, and the Research Ethics Review Committee (ERC) of the World Health Organisation (WHO) recommends: 'Explain the procedures and any medical terminology in simple language. Focus on what is expected of the child'.[4] In this instance, a child cannot be expected to completely understand medical science or terminology; rather, a researcher should convey the aspects of the treatment that are estimated to be essential at the time of taking the informed consent of adults, for example, the progression of the disease, the potential risks and expectations. A child's informed consent is valid if and only if the researcher has made the child understand all the requisite essential details to comply with research ethics.

AI can help medical personnel by collaborating with previously conducted research about the capacity of children [12]. For example, if AI predicts the probability of a certain sequela, then it should additionally paraphrase the concept of sequela and probability in plain language for the lower limit of the child's comprehension abilities. Further, medical professionals should not neglect their duty to communicate with the child's parents, who are asked to tender consent by proxy [50]. There is a two-step coarsening $E_{Professional} \rightsquigarrow E_{Parents} \rightsquigarrow E_{Child}$ and each step can be validated through the rationale of impracticability, the lack of knowledge and understanding.

4.3 Coarse justification for perspective adjustment

4.3.1 Basic idea

Coarse equation usually depends on the question of from which standpoint the regulator assesses an activity. For example, in debating global warming, some people may counter with the argument that this climate change is trivial because humanity will one day die out on the timescale of earth's history. Such trivialisation is also found among the people who live only for the present without concern for the future. It is hence important for CE to answer the question of the perspective to be adopted at the time of evaluation. The lens must neither be too wide as in the case of earth's history nor too narrow as in the case of living for the moment. In a reflexive manner, coarsening can be endorsed as an evaluation method using the rationale that the original method adopts too wide or too narrow a perspective.

4.3.2 Definition

An original evaluation E_{i} adjudicated the relationship between two objects o_{i} and o_{j} as $o_{\mathrm{i}} < o_{\mathrm{j}}$, but a coarsened E_{j}

[3] This moral postulate is classical. An alternative formulation is already found in Digest 50.17.185 of the Justinian Code of Roman law: 'There is no obligation to do anything which is impossible' (translated by Watson et al. [51] p. 482, in Latin: *Inpossibilium nulla obligatio est*).

[4] This formular is published by *WHO, Informed Assent Form Template for Children/Minors*, <https://www.dcu.ie/sites/default/files/research_support/who_informedassent.doc> accessed 05 February 2021, p. 4.

equated the assessment as $o_i \approx o_j$. In this case, if and only if E_i is not disputed by the stakeholders and coarsened into E_j merely for the purpose of perspective adjustment, the evaluator can justify the coarsening simply through the rationale of a shift in perspective. This justification is labelled *coarse justification for perspective adjustment*.

A well-known Chinese narrative can facilitate the understanding of this type of justification. In ancient China, King Hui of Wei consulted with Mencius, one of the most influential Chinese Confucian thinkers, about the reason why his Kingdom could not attract immigrants although his governance was better than the rule of other kings. Mencius did not answer directly; instead, he queried:

> Your Majesty loves war; allow me to take an illustration from war. The soldiers move forward at the sound of the drum; and when the edges of their weapons have been crossed, on one side, they throw away their buff coats, trail their weapons behind them, and run. Some run a hundred paces and then stop; some run fifty paces and stop. What would you think if these, because they had run but fifty paces, should laugh at those who ran a hundred paces? [34]

King Hui answered there was no difference between them. Receiving this response, Mencius explained that the distinction between King Hui and the other kings was merely such a difference from the viewpoint of ideal Confucian politics. In this case, the argument offered by Mencius can be compared with coarse justification through the shift in perspective. Mencius did not appraise King Hui's administration in detail; instead, he offered the rough analogy of two cowardly soldiers to contend that King Hui's reign could be equated with the governance of the other monarchs from the perspective of those who were ruled. Here King Hui and Mencius did not dispute whether sovereigns should perform good governance. Mencius encouraged the King to merely adjust his perspective from that of a ruler to that of the ruled, and Mencius's rough explanation was justified even though his analogy was ambiguous from the strict viewpoint of modern political sciences.

The process of discovering a good perspective should normally be contemplated: at what scale should AI evaluate human activities? This question remains open in the present study, however, and the easier question of the process of discovering a bad perspective is answered instead. A perspective that appears to be selected ad-hoc to avoid criticisms is a clear answer to that question. The trite argument against veganism that lions eat meat may be cited at this juncture. Biehl [7] points out that lions do not denote an appropriate benchmark for the determination of an ethically sound decision. This anti-veganism argument can easily be defeated using CE, because the behaviour of lions is not usually contemplated in human dietary decision-making. Thus, the reference to lions in the context of veganism is clearly ad-hoc. AI should not excuse ad hocism either.

Another problematic perspective change involves the shift from the minority to the majority viewpoint, for example, when racial bias is possible [26, 48]. Xiang et al. [54] emphasise that compatibility with the law is required for equitable AI predictions; however, fairness is not achieved if AI always sides with the majority perspective and hides the people who are discriminated against. Thus, the principle of justice is imperative for algorithmic decision-making: Holstein et al. [25] call it *algorithmic fairness*. Since fairness relates to the prohibition of biases against certain protected characteristics [8], it leads to the conclusion that the majority and minority group should be treated equally. If a minority group performs a religious ritual that differs from the ceremonies performed by the majority, AI should not cater to the majority and assess the minority rite as socially useless, because the freedom of religion is a fundamental human right. The concept of explainable does not imply a decision that is comprehended and accepted by the majority.

5 Application

Given the formality of trade-offs, this section discusses their application in automated driving as an example of XAI. GenEth is a programme that uses ILP to extract ethical principles, and when given cases are input, it extracts an ethical principle that is consistent with these cases [1]. The basic idea of GenEth is as follows: first, given a choice of two actions $a1$ and $a2$, these two actions are evaluated against a specific obligation d; second, an ethical value is assigned to each action, and the user decides in advance whether $a1$ or $a2$ should be preferred; lastly, after the values of $a1$ and $a2$ and the preferences are input for all assumed cases, GenEth extracts general principles that are consistent with those preferences (i.e. the correct choice can be made for all cases).

Anderson et al. [1] gave two options in the context of automated driving, that is, AI can either switch to automatic driving mode or not, and they took five duties for AI, namely, (1) avoiding a collision, (2) staying in lane, (3) respecting the driver's autonomy, (4) keeping within the speed limit and (5) avoiding imminent harm to people. Anderson et al. [1] simplified the evaluation value using only integer numbers; for example, by assigning 1 if the act satisfies the obligation and − 2 if it violates the obligation seriously. They defined Δd as the difference between the values of $a1$ and $a2$ (i.e. $d_{a1} - d_{a2}$) when viewed from a particular duty d (e.g. if $d_{a1} = +1$ and $d_{a2} = -1$, then $\Delta d = +2$). GenEth analyses the relationship between these two options and the five obligations in several hypothetical cases and extracts a general principle that is consistent with them.

Table 4 An example for automated driving evaluation by DNN

Actions	Avoiding a collision	Staying in lane	Respecting the driver's autonomy	Keeping with the speed limit	Avoiding imminent harm to people
$a1$: do not take control	− 1.01	0.61	0.64	0.67	− 1.38
$a2$: take control	1.18	− 0.66	− 0.49	− 0.85	1.27
Δd	− 2.19	1.27	1.13	1.52	− 2.65

Table 5 An example for inappropriate simplification by rounding off

Actions	Avoiding a collision	Staying in lane	Respecting the driver's autonomy	Keeping with the speed limit	Avoiding imminent harm to people
$a1$: do not take control	− 1	1	1	1	− 1
$a2$: take control	1	− 1	0	− 1	1
Δd	− 2	2	1	2	− 2

If Δd takes all real values (e.g. $\Delta d = 0.0001$), then GenEth may be able to trace them. However, in practice, the function to represent such small decimals is not implemented in the version that Anderson et al. [1] introduced in their paper (see their Fig. 3), and they dealt only with integer values. Of course, Anderson et al. [1] also avoided expressing an obligation in terms of only three values, such as +1 for fulfilment, −1 for violation and 0 for irrelevance; for example, according to the duty of avoiding imminent harm to persons, taking control by AI is +1 in the minor personal injury but +2 in the major one. However, this difference still involves a simplification because it is clearly simplistic to say that the value of the latter is twice that of the former. Since GenEth establishes norms from the input integer value set, the simplification of input values has a direct impact on the norms.

This is where CE contributes to the solution. Consider the following fictional case: a group of AI developers devised a deep-learning model for evaluating the five obligations that an automated driver should fulfil, as in the case of Anderson et al. [1]; they applied it to a pedestrian accident case and generated Table 4.

As the developers were unable to explain the exact meaning of each value in this table, they decided to enter the values in GenEth and make a knowledge-based representation. For entering these values into GenEth, they rounded off each decimal of $a1$ and $a2$ to the nearest integer number and recalculated each Δd. For example, they guessed that the following formula would be valid with regard to the obligation of avoiding a collision.

$$-1.01 < {}^{\text{DNN}}_{\text{Better}} -1 \rightsquigarrow -1.01 \approx {}^{\text{GenEth}}_{\text{Equal}} -1$$

Thus, they obtained the values shown in Table 5.

Here, the developers argued that the following formula would hold:

$$a1(-1.01, 0.61, 0.64, 0.67, -1.38) < {}^{\text{DNN}}_{\text{Better}}\ a2(1.18, -0.66, -0.49, -0.85, 1.27)$$

$$\rightsquigarrow a1(-1, 1, 1, 1, -1) < {}^{\text{GenEth}}_{\text{Better}}\ a2(1, -1, 0, -1, 1)$$

Is the trade-off between accuracy and interpretability formally reasonable in this case? From the perspective of CE, this formula needs to meet the two requirements of adequately high coverage and order-preservation. The first requirement requests that as many of the obligations as possible in Table 4 should be covered by Table 5. Since there are only five obligations, it can be said that the coverage of all obligations is adequately high. Table 5 covers all of them; and thus, the first requirement is satisfied. Second, from the requirement of order-preservation, the value order of Table 5, which is roughened, must match the original value order of Table 4. However, Table 5 fails to satisfy this requirement because it seems reasonable to take control in Table 4 and not in Table 5, for the relation between the total values of $a1$ and $a2$ is $a1 = -0.47 < a2 = 0.45$ in Table 4 but $a1 = 1 > a2 = 0$ in Table 5; in other words, there is a reversal that is forbidden by the requirement of order-preservation. Thus, simplification by rounding off is an incorrect procedure from the perspective of CE—at least for this case.

As more and more cases are entered into GenEth, it becomes more frequent that rounding off no longer preserves the original value order; therefore, there is a need to either develop another good method for rounding off or avoid trying to pursue explainability by rounding off the numbers predicted by DNN in the first place. For example, multiplying each value by 100 may seem like a good

idea, but substantial changes in the figures may give rise to other risks. What is important here is that this discussion is completed by a formality check alone without any complex ethical arguments, such as about the incompatibilities in the ways different kinds of cars function and interact with each other [37], and this formality check can offer suggestions for the improvement of transparent models.

6 Conclusion

AI should be explainable, but the concept of XAI relates to the problem that human interpretability entails a trade-off with the accuracy of AI prediction. Hence, a branch of ethics is required to tackle questions pertaining to the circumstances in which accuracy can be sacrificed in favour of human interpretability. This study proposes the conception of CE, which does not assume the asymptotic approach to the ideal, unlike Kantian ethics. Instead, CE permits regulators, judges and other actors to mitigate their normative activities. This idea cannot determine which trade-off pattern is the best, but it can tell AI developers what trade-offs should be prohibited for breach of formality.

There are at least two minimal requirements that differentiate inadequate and adequate coarsening: an adequately high coverage and order-preservation. Adequately high coverage signifies that a rough evaluation should sufficiently encompass objects assessed using the original method; order-preservation is a task that helps to sustain the value order of the original evaluation during its coarsening. For example, if a model of XAI tries to explain a symptom by discarding most of the multiple features that an original deep-learning model focuses on, then this XAI model does not meet the first requirement. Likewise, when financial AI assesses customers for loan, the priority of each customer should not be reversed between the original deep-learning model and XAI; otherwise, the second requirement would not be satisfied.

Meeting these two requirements should evince the conditions in which normative coarsening may be justified. This study posited two arguments for the validation of coarsening: impracticability and perspective adjustment. First, an evaluator may coarsen an impracticable prediction to a practicable option, and AI can use this method to assure human interpretability. To cite an example, AI can sacrifice accuracy and roughen a prediction in favour of another that is more understandable or performable for human agents if AI estimates that the prediction would not be comprehensible or performable for its users. Second, coarsening may be justified for perspective adjustment to allow an evaluator to shift from too wide or too narrow a viewpoint to one that fits a given situation. It would be difficult to find the best perspective at this juncture; hence, this study evinced two examples of inadequately changing standpoints: the ad-hoc view and majority opinions. XAI should not use the relativity of perspectives either as an excuse or as a means of discrimination .

The above discussion can also contribute to the development of XAI. Using GenEth as an example, this study suggests that GenEth should be able to handle not only integer but also a number with a decimal point. This suggestion may also apply to other models that have a limited set of available values or use broad concepts for inference. In conclusion, CE would be pivotal in facilitating inter-connections between interdisciplinary research initiatives on XAI and would offer directions for the handling of the trade-off between computable accuracy and human interpretability.

Funding No funding.

Declarations

Conflict of interest No conflicts of interest.

Data availability statement No specific data, material or code.

References

1. Anderson, M., Anderson, S.L.: GenEth: a general ethical dilemma analyzer. Paladyn. J. Behav. Robot. **9**, 337–357 (2018). https://doi.org/10.1515/pjbr-2018-0024
2. Arrieta, A.B., Díaz-Rodríguez, N., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform. Fus. **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
3. Baer, B.R., Gilbert, D.E., Wells, M.T.: Fairness criteria through the lens of directed acyclic graphs. In: Dubber, M.D., et al. (eds.) The Oxford Handbook of Ethics of AI, pp. 493–520. Oxford University Press, New York (2020)
4. Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., Šabanović, S.: Human-Robot Interaction: An Introduction. Cambridge University Press, Cambridge (2020)

5. Barfield, W., Barfield, J.: An introduction to law and algorithms. In: Barfield, W. (ed.) The Cambridge Handbook of the Law of Algorithms. Cambridge University Press, Cambridges (2020)
6. Bethel, C.L., Henkel, Z., Baugus, K.: Conducting studies in human-robot interaction. In: Jost, C., et al. (eds.) Human-Robot Interaction. Springer Series on Bio-and Neurosystems 12, pp. 91–124. Springer, Cham (2020)
7. Biehl, L.: Lions eat meat too: when lions dictate what's morally right. The Animalist. (2019) https://the-animalist.ch/en/arguments-veganism/. Accessed 11 Feb 2021
8. Boddington, P.: Normative modes: codes and standards. In: Dubber, M.D., et al. (eds.) The Oxford Handbook of Ethics of AI, pp. 125–140. Oxford University Press, New York (2020)
9. Chrisley, R.: A human-centered approach to AI ethics. In: Dubber, M.D., et al. (eds.) The Oxford Handbook of Ethics of AI, pp. 463–474. Oxford University Press, New York (2020)
10. Coeckelbergh, M.: AI Ethics. The MIT Press, Cambridge (2020)
11. Contissa, G., Lagioia, F., Sartor, G.: The ethical knob: ethically-customisable automated vehicles and the law. Artif. Intell. Law **25**, 365–378 (2017). https://doi.org/10.1007/s10506-017-9211-z
12. Daly, A.: Assessing children's capacity. Int J Child Rights **28**, 471–499 (2020). https://doi.org/10.1163/15718182-02803011
13. Dautenhahn, K.: Socially intelligent robots: dimensions of human-robot interaction. Phil. Trans. R. Soc. B **362**, 679–704 (2007). https://doi.org/10.1098/rstb.2006.2004
14. Dunham, J., Grant, I.H., Watson, S.: Idealism: The History of a Philosophy. Routledge, London (2010)
15. Fox, M., Long, D., Magazzeni, D.: Explainable planning. (2017) Available at: arxiv:1709.10256v1
16. Friedman, M.: Regulative and constitutive. Southern J Philos **30**(S1), 73–102 (1992). https://doi.org/10.1111/j.2041-6962.1992.tb00658.x
17. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models. (2018) Available at: arxiv:1802.01933v3
18. Gunning, D., Aha, D.W.: DARPA's explainable artificial intelligence (XAI) program. AI Mag. **40**(2), 44–58 (2019). https://doi.org/10.1609/aimag.v40i2.2850
19. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: XAI-Explainable artificial intelligence. Sci Robot **4**(37), eaay7120 (2019). https://doi.org/10.1126/scirobotics.aay7120
20. Hall, P.: On the art and science of machine learning explanations. In: 2019 KDD XAI workshop. (2018) Available at: arxiv:1810.02909v4
21. Hamon, R., Junklewitz, H., Sanchez, I.: Robustness and Explainability of Artificial Intelligence. In: EUR 30040 EN, Publications Office of the European Union, Luxembourg (2020). https://doi.org/10.2760/57493
22. Haraway, D.: Simians, cyborgs, and women. Routledge, London (1991)
23. Hiller, A., Woodall, T.: Everything flows: a pragmatists perspective of trade-offs and value in ethical consumption. J Bus Ethics **157**, 893–912 (2019). https://doi.org/10.1007/s10551-018-3956-5
24. Hobbes, T.: Leviathan. (1651) Project Gutenberg: https://www.gutenberg.org/ebooks/3207
25. Holstein, T., Dodig-Crnkovic, G., Pelliccione, P.: Steps towards real-world ethics for self-driving cars: beyond the trolley problem. In: Thompson, S.J. (ed.) Machine Law, Ethics, and Morality in the Age of Artificial Intelligence, pp. 85–107. IGI Global, Pennsylvania (2021)
26. Hong, J.W., Williams, D.: Racism, responsibility and autonomy in HCI: testing perceptions of an AI agent. Comput. Hum. Behav. **100**, 79–84 (2019). https://doi.org/10.1016/j.chb.2019.06.012
27. Horty, J.F.: Agency and Deontic Logic. Oxford University Press, New York (2001)
28. Jasanoff, S.: The Ethics of Invention: Technology and the Human Future. W. W. Norton, New York (2016)
29. Kant, I.: The Metaphysics of Morals. In: Denis, L. (Gregor, M. translator) (ed.) Cambridge University Press, Cambridge (2017)
30. Kroll, J.A.: Accountability in computer systems. In: Dubber, M.D., et al. (eds.) The Oxford Handbook of Ethics of AI, pp. 181–196. Oxford University Press, New York (2020)
31. Lee, M.S.A., Floridi, L., Singh, J.: Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00067-y
32. Liao, B., Anderson, M., Anderson, S.L.: Representation, justification, and explanation in a value-driven agent: an argumentation-based approach. AI Ethics **1**, 5–19 (2021). https://doi.org/10.1007/s43681-020-00001-8
33. Lundberg, S., Lee, SI.: A unified approach to interpreting model predictions. In: The 31st Conference on Neural Information Processing Systems. (2017) Available at: arxiv:1705.07874v2
34. Mencius: The sayings of Mencius. In: Epiphanius, W. et al. (eds.) Chinese Literature: Comprising the Analects of Confucius, the Sayings of Mencius, the Shi-King, the Travels of Fâ-Hien, and the Sorrows of Han. (1900) Project Gutenberg: https://www.gutenberg.org/ebooks/10056
35. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. (2018) Available at: arxiv:1706.07269
36. Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. (2021) Available at: https://christophm.github.io/interpretable-ml-book/
37. Nyholm, S.: Humans and Robots: Ethics, Agency, and Anthropomorphism, Illustrated Rowman & Littlefield Publishers, Lanham (2020)
38. Rabold, J., Schwalbe, G., Schmid, U.: Expressive Explanations of DNNs by Combining Concept Analysis with ILP. In: Schmid, U., et al. (eds.) KI 2020: advances in artificial intelligence. KI 2020. Lecture Notes in Computer Science, vol. 12325, pp. 148–162. Springer, Cham (2020)
39. Raso, F.A., Hilligoss, H., Krishnamurthy V., Bavitz, C., Kim, L.: Artificial intelligence and human rights: Opportunities and risks (September 25, 2018). Berkman Klein Center Research Publication No. 2018–6 (2018). https://doi.org/10.2139/ssrn.3259344
40. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. KDD' 16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778
41. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
42. Scarborough, P., Appleby, P.N., Mizdrak, A., Briggs, A.D.M., Travis, R.C., Bradbury, K.E., Key, T.J.: Dietary greenhouse gas emissions of meat-eaters, fish-eaters, vegetarians and vegans in the UK. Clim. Change **125**, 179–192 (2014). https://doi.org/10.1007/s10584-014-1169-1
43. Schwartz, D.: Probabilism reconsidered: deference to experts, types of uncertainty, and medicines. J. Hist. Ideas **75**(3), 373–393 (2014)
44. Setchi, R., Dehkordi, M.B., Khan, J.S.: Explainable robotics in human-robot interactions. Proc. Comput. Sci. **176**, 3057–3066 (2020). https://doi.org/10.1016/j.procs.2020.09.198
45. Sidobre, D., Broquère, X., Mainprice, J., Burattini, E., Finzi, A., Rossi, S., Staffa, M.: Human-robot interaction. In: Siciliano, B. (ed.) Advanced Bimanual Manipulation, pp. 123–172. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-29041-1_3
46. Takeda, M., Hirata, Y., Weng, Y.H., Katayama, T., Mizuta, Y., Koujina, A.: Verbal guidance for sit-to-stand

support system. Robomech J **7**, 8 (2020). https://doi.org/10.1186/s40648-020-00156-3
47. The High-Level Expert Group on Artificial Intelligence: Assessment List for Trustworthy Artificial Intelligence. (2020) Available at: https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1
48. Tian, J., Xie, H., Hu, S., Liu, J.: Multidimensional face representation in a deep convolutional neural network reveals the mechanism underlying AI racism. Front. Comput. Neurosci. **15**, 620281 (2021). https://doi.org/10.3389/fncom.2021.620281
49. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): towards medical XAI. IEEE Trans. Neural Netw. Learn. Syst. (2020). https://doi.org/10.1109/TNNLS.2020.3027314
50. Varadan, S.: The role of parents in the proxy informed consent process in medical research involving children. Int. J. Child. Rights **28**(3), 521–546 (2020). https://doi.org/10.1163/15718182-02803009
51. Watson, A. (ed.): The Digest of Justinian, vol. 4. University of Pennsylvania Press, Philadelphia (1998)
52. Weng, Y.H., Izumo, T.: Natural law and its implications for AI governance. Delphi **2**(3), 122–128 (2019). https://doi.org/10.21552/delphi/2019/3/5
53. Winikoff, M.: Towards trusting autonomous systems. In: Seghrouchni, A.E.F., et al. (eds.) Engineering Multi-Agent Systems, pp. 3–20. Springer, Cham (2018)
54. Xiang, A., Raji I.D.: On the legal compatibility of fairness definitions. In: Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems. (2019) Available at: arxiv:1912.00761v1
55. Yeung, K., Howes, A., Pogrebna, G.: AI governance by human rights-centered design, deliberation, and oversight. In: Dubber, M.D., et al. (eds.) The Oxford Handbook of Ethics of AI, pp. 77–106. Oxford University Press, New York (2020)