Encoding Social & Ethical Values in Autonomous Navigation: Philosophies Behind an Interactive Online Demonstration

Yun Tang
SATM, Cranfield University
Cranfield, Bedfordshire, UK
yun.tang@cranfield.ac.uk

Luke Moffat
Sociology, Lancaster University
Lancaster, UK
l.moffat1@lancaster.ac.uk

Weisi Guo*
SATM, Cranfield University
Cranfield, Bedfordshire, UK
weisi.guo@cranfield.ac.uk

Corinne May-Chahal
Sociology, Lancaster University
Lancaster, UK
c.may-chahal@lancaster.ac.uk

Joe Deville
Management, Lancaster University
Lancaster, UK
j.deville@lancaster.ac.uk

Antonios Tsourdos
SATM, Cranfield University
Cranfield, Bedfordshire, UK
a.tsourdos@cranfield.ac.uk

Abstract

Autonomous Systems (ASs) interacting with human societies raises complex social & ethical challenges. This paper argues that one way of scaffolding human trust in ASs is through the encoding of ethical, legal and social impact (ELSI) considerations in the ASs' decision-making processes. Existing ELSI-encoding efforts often focus on the implementation of rule-based and risk-based approaches, leaving key questions unanswered - what are the relationships between ELSI-encoding software logic in ASs and human ethical practises; what ethical approaches cannot be easily translated into software rules and numeric risks; and what are the implications of this for ethical AS?

To answer these questions, we review and discuss different ELSI-encoding approaches in ASs from a new perspective, i.e., their relationships with classic human ethics philosophies. We also explore the feasibility of large language models (LLMs)-based ELSI-encoding practices in overcoming the limitations of rule-based and risk-based approaches and the associated challenges. To foster understanding, facilitate knowledge exchange and inspire discussion among cross-disciplinary research communities, we build and publish the first online interactive playground demonstrating different ELSI-encoding approaches on the same AS decision-making process. We welcome feedback and contributions in making this platform truly beneficial to trustworthy autonomous system research communities.

CCS Concepts

• Applied computing → Sociology; **• Social and professional topics → User characteristics**; **• Human-centered computing → Empirical studies in HCI**.

Keywords

Trustworthy Autonomous Systems, ELSI-encoding, Large Language Model, Demonstration

ACM Reference Format:
Yun Tang, Luke Moffat, Weisi Guo, Corinne May-Chahal, Joe Deville, and Antonios Tsourdos. 2024. Encoding Social & Ethical Values in Autonomous Navigation: Philosophies Behind an Interactive Online Demonstration. In *Second International Symposium on Trustworthy Autonomous Systems (TAS '24), September 16–18, 2024, Austin, TX, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3686038.3686044

1 Introduction

Autonomous systems (ASs) can accelerate many complex tasks such as mission planning and route navigation, often with no or few human interventions [3]. One sector that has significantly benefited from autonomous systems is transportation, particularly through autonomous vehicles [29], while other sectors are also reaping the rewards of enhanced autonomy [19, 32, 33].

1.1 Relationship between Trust and Social Value Encoding

Despite many potential benefits, ASs have yet to proliferate in our daily lives and one of the main reasons is the lack of human trust in its connected technologies [22], with examples such as robotic encounters [17] and anxiety around human-in-the-loop autonomy [29]. Gaining people's trust is a difficult and complex task because it is about more than simply a supply of sufficient information.

A way to tackle this is to acknowledge the central position of stakeholders and their values, exploring how these might be incorporated into AS design practices. Researchers in recent years have started to propose methods to encode ethical, legal OR social norms into the design of autonomous systems. However, they are scattered across different application domains and we do not have a common ground for stakeholders from different communities to compare and discuss the broader prospect of ELSI-encoding approaches. We propose that enhancing the communicative pathway between (1) ELSI principles of different stakeholders and (2) autonomous decision-making will improve the efficacy of the system by enhancing multi-faceted representations of trust between the trustee (AS) and the trustor (human end-user). One way of approaching trust

*Corresponding Author

TAS '24, September 16–18, 2024, Austin, TX, USA

ACM ISBN 979-8-4007-0989-0/24/09
https://doi.org/10.1145/3686038.3686044

is through four interrelated pillars [27] of: *ability*, *benevolence*, *integrity* of the trustee, and the trustor's *propensity* to trust. In this paper, we are focusing particularly on:

- the *benevolence* (e.g., the extent to which the AS trustee is seen as having a genuine concern for the welfare of the trustor end-user, beyond any contractual or duty-bound obligations) and
- the *integrity* (e.g., the trustor's belief that the trustee consistently follows a set of principles or values that are agreeable or acceptable to the trustor).

One of the motivations for this paper and the demos discussed in it is to further understand and make contributions to responsible AI, where responsibility is not just taken as the liability of users or regulators, but as a key social and ethical value through AS design processes. Making AI responsible is not just about fulfilling criteria, but involves complex sets of agreements, including agreements to trust, between AS designers, engineers, manufacturers, regulators, users, and communities.

1.2 Main Contribution and Organisation

We review and discuss different ethics-encoding approaches in ASs from a new perspective, i.e., their relationships with ethical philosophies. We also explore the feasibility of large language models (LLMs)-based ELSI-encoding practices in overcoming the limitations of rule-based and risk-based approaches. Based on the above, we design and publish the first online interactive playground allowing users from different communities to engage in the ELSI-encoding designing phase of AS through role-playing different stakeholders such as AS developers and users. This will foster understanding and facilitate knowledge exchange by comparing AS performances with different social value-encoding approaches of one common decision-making process. Note that while there are works on enhancing explainability [16] or evaluating the trustworthiness [39] of autonomous systems, we focus on the practical approaches to encoding ELSI principles, particularly social and ethical values, in the decision-making processes of autonomous systems.

The structure of our paper is as follows. We first review in Section 2 different ethics-encoding approaches for autonomous systems across application domains. In Section 3, we discuss the linkage between the ELSI-encoding practices in ASs with different traditional ethical philosophies practised by humans. Backed by this, we construct in Section 4 a website-based interactive playground, showcasing different social value-encoding approaches for a rudimentary navigation planning process for autonomous vehicles. In Section 5, we discuss the questions raised by comparing the ELSI-encoding approaches.

2 Encoding ELSI Principles in ASs

Existing methodologies for encoding ethical principles into the decision-making processes of autonomous systems can be divided into four categories. Before summarising these, it is worth first pointing out some of the complexities involved with integrating social values with AI design practices.

The last 10 years have seen an explosion of academic research and publications on the importance of ethics in autonomous technologies. This has resulted in countless calls for ethical frameworks, guidelines, and principles to help manage the impacts and uncertainties generated by AS design. While principles are important, as demonstrated in the discussions below, it is also important to remember that part of the value of ethics is in raising questions, confronting complexity, and pointing out the often insoluble dilemmas associated with how human beings interact with others. Social values are contingent upon relations, including relations with technology. As such, while we show in this paper the contributions that an ELSI-informed approach can make to AS design, we do not claim to have solved all the problems. In section 5, we point out some of the ways in which the presented demos generate further ethical questions. In short, just as we should not expect AS to be a silver bullet to all social problems, we should also not expect ethics to be a silver bullet to AS impacts.

2.1 Hard-coded "Common Sense"

The most rudimentary and intuitive way of encoding ethical principles is through direct translation of the "customs and practices" associated with ELSI into the parameters or mathematical equations in the decision-making processes of autonomous systems. For example, for self-driving cars, Shalev et al. [35] proposed a Responsibility-Sensitive Safety model to formalize the "common sense" driving rules (e.g., *do not tailgate*) using mathematical equations and parameters. Specifically, the rule "*do not tailgate*" can be translated into a safe car-following distance considering parameters such as the maximum comfortable deceleration rate and average human response time. Similarly, Parnell et al. [32] define the rules for an autonomous assistive dressing robot in a domain-specific language called SLEEC, which is proposed by [42] for specifying social, legal, ethical, empathetic and cultural constraints for autonomous systems. Specifically, the rule "*the assistive robot must call support regardless of user consent if it detects the user has fallen*" can be expressed using the SLEEC language as "***rule_start*** *rule_id* ***when*** *UserFallen* ***then*** *SupportCalled* ***unless not*** *assentToSupportCalls* ***rule_end***" where *UserFallen* and *SupportCalled* are action events detected during the operation of the autonomous robot and *assentToSupportCalls* are measurements (e.g., user's assent as a boolean can be detected via facial recognition). Once the rules are determined, the events (e.g., *UserFallen*) and measurements are mapped to the actionable verbs (e.g., detect), physical objects (e.g., human user), and object-related processes (e.g., detecting user and his location). In such a way, the high-level behaviour strategies of the autonomous systems can be regulated by pre-defined ethical rules.

2.2 ELSI Risk Estimation and Minimization

As the situation faced by autonomous systems gets complex and the number of contributing factors for decision-making increases, implementing ELSI principles for every situation considering every combination of factor values becomes impractical. "Common sense" customs are translated into rules as a facsimile of social interaction. The practical application of these rules, however, lacks the reflexivity necessary for continuous adaptation to changing contexts and

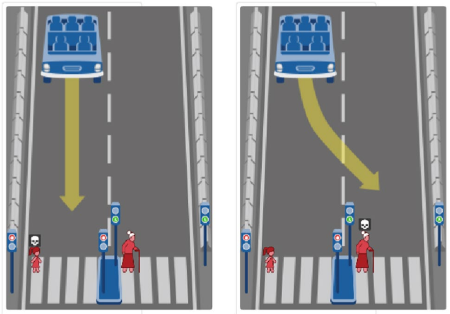


Figure 1: The ethical dilemma discussed in [25] where the autonomous vehicle needs to choose between hitting a young girl (illegally passing the pedestrian crossing) and an old lady (rightfully passing the pedestrian crossing). Autonomous vehicles in [25] should go straight instead of swerving due to lower estimated risk considering all the ELSI factors.

situations. Rules alone lack reflexivity because they are prescriptive. The challenge then becomes how to add to these rules a level of agility in the face of uncertainty.

A more scalable and generalizable risk-based approach has been proposed to address this. For example, Geisslinger et al. [14] proposed an ethical trajectory planning framework for autonomous vehicles adopting the "ethics of risk", which is quantified by the estimated collision probability and harm severity for every possible trajectory and the trajectory with the least numerical ethical risk is selected. Later, Liu et al. [25] extended the risk-based framework and demonstrated it through a moral dilemma scenario shown in Fig. 1. The main contributing factors (i.e., species, harm, traffic law, number and age) are determined by literature review and their normalized weights are calibrated with public opinions through questionnaires. Finally, the risk degree for each action is calculated as a weighted sum of the contributing factors (e.g., 0.8 for going straight and 1.2 for swerving) and the action "*going straight*" is chosen by the autonomous vehicle as it is less morally risky.

2.3 LLM-assisted ELSI Reasoning

Hard-coded rule-based or risk estimation-based approaches are sufficient when the mapping between ELSI rules and situations (faced by autonomous systems) or the list of contributing factors is explicit. However, ELSI principles can be abstract, situations can be countless, risk factors can be hidden, and human feedback or commands can be unpredictable. Recent work [4] shows that how abstract rules are broken and interpreted are subjective to personal preferences, further challenging the hard-coded (and fixed) rule-based or risk-based approaches. To handle unknown situations or reason about abstract ELSI principles, large language models (LLMs) can be utilized.

LLMs such as GPT [30], Claude [2] and Gemini [15] have been widely adopted as decision-making assistants in autonomous systems such as robotic arms [18] and autonomous vehicles [26] due to their promising memorization, comprehension and reason capabilities [10, 43]. The training data of LLMs is extracted from an enormous amount of human-written articles, posts, or discussions from the internet, and thus inevitably contains the knowledge of concepts, human preferences, application examples or discussions around ELSI principles. The wealth of ELSI-related knowledge baked into the LLM model parameters during training can be systematically evaluated using the methodology proposed by Tang et al. in [37]. As a result, LLM can be specially asked to infer the relevant ELSI principles given the described decision-making context and consider them during decision generation.

For example, Luu et al. [26] utilize an LLM to analyze latent events (e.g., children suddenly rushing across the road from behind the school bus) in autonomous driving scenarios and generate safe actions (e.g., slowing down when approaching the school bus) based on the inferred latent objects (e.g., children). Constant et al. [11] propose an LLM-assisted ETLC (Extract, Transform, Load, Compute) framework to encode the traffic laws into the decision-making processes of autonomous vehicles. Specifically, LLM extracts the legal requirements (e.g., "*an autonomous vehicle may be operated on the public road for testing purposes by a driver who possesses the proper class of license for the type of vehicle being operated if all the following requirements are met ...*" from California Vehicle Code article 38750 clause b [21]) and transforms into legal decision paths modelled by directed graphs containing three types of nodes, *consequence* node (*the autonomous vehicle may be operated*), the *criteria* node (*the operation is on the public road for testing purpose and with a qualified driver ...*) and the *evidence* node (the observed evidence supporting a criterion node). Such a decision paths graph is then implemented directly for decision-making by the autonomous vehicle.

2.4 Human Supervision-enabling Interface

Developing a specialized interaction interface between human and autonomous systems and enabling human users to monitor, understand and control the decision-making processes of autonomous systems at runtime is another commonly adopted approach for human trust. For example, Wilson et al. [41] discovered that humans find it challenging to predict and understand the complex, emergent behaviours of robotic swarms and called for the development of user interfaces that help humans monitor and control the swarms effectively without being overwhelmed. Similarly in [20], after recognizing that the online recommender systems, integral to Netflix and Spotify, typically prioritize immediate user preferences (first-order desire [13]) but may neglect deeper, reflective desires (second-order desire [13]), Krook and Blockx also call for a mechanism to allow users to reflect their desires and adjust their preferences to promote human user autonomy and reduce the risk of manipulation. In addition to the configuration or command interface which requires active human engagement, sensors can also be used for unobtrusive human feedback collection. In the TEACHING project [5], Bacciu et al. developed a service employing unobtrusive wearable and environment sensors to collect human psychological, emotional and cognitive states used by autonomous vehicles to adapt their behaviour during operation.

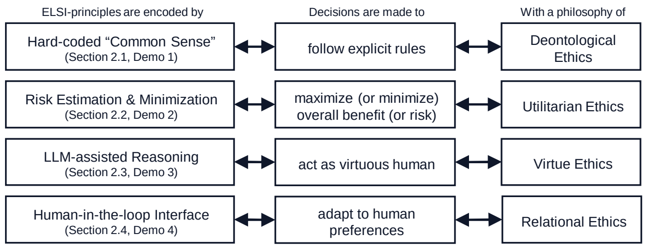


Figure 2: ELSI-encoding approaches in autonomous systems and the corresponding philosophies of ethics behind.

3 A New Perspective on ELSI-encoding Approaches

According to the survey results of the ethical dilemma scenario in Fig. 1, about 86% of the respondents chose to go straight and hit the young girl [25]. However, there were still 14% of the respondents choosing to swerve toward the old lady and we cannot mark the choice as incorrect simply because it deviates from the majority. For example, we might get the opposite survey results if the survey is done in another population according to the Cultural Relativism theory [38]. It is widely recognized that individuals apply their ethical theories to distinguish between what is good and bad, or right and wrong. This view of human decision-making resonates with some of the classical ethical theories rooted in Philosophy [34]. In what follows, we summarise some of the dominant models of ethics that find their way into AS research, as well as a less discussed view of ethics that has informed the creation of the demo.

Deontology characterises the morality of actions according to their alignment with principles. Its influence in AS research derives largely from Kantian ethics, in which a person is morally correct to the extent that they follow rules supplied to them by reason. Deontological ethics is therefore inherently rationalist. While it cannot explain precisely what characterises a given action as good or bad since it is not the action itself but its adherence to principles that gives it moral significance, nevertheless deontology insists a person's capacity for ethical behaviour is rooted in their capacity for rational thought. In this case, moral principles are unchanging and transcend contextual circumstances. For instance, if lying could prevent a crime, a deontologist such as Kant would still consider it wrong because lying is inherently unethical. It violates the moral rule of always being truthful. Vitally, deontological ethics denies that the moral character of an action can be derived from its potential consequences, because these consequences cannot be guaranteed beforehand.

Consequentialism, also known as **Utilitarianism**, bases morality on the outcomes of actions, rather than the actions themselves. For a utilitarian, such as Mill, an action is right if it results in the greatest good for the greatest number of people, with the least amount of harm. If a doctor has five patients who need organ transplants and one healthy person whose organs could save all five, consequentialism might argue that it would be morally permissible to sacrifice the one to save the five, as this results in the greatest overall good. One criticism of consequentialism is that it can be used to justify actions we might consider impermissible or unacceptable, because of its focus on the outcome. One could say that reducing the population might result in a better or more sustainable life for those who remain but says nothing of what measures might be taken to achieve that end.

Virtue Ethics emphasizes an individual's character and the virtues that constitute a good person rather than specific actions or consequences. It is more about developing good character traits and being a morally good person. If a person finds a wallet full of money, a virtue ethicist would focus on what a virtuous person would do, which likely would be to return the wallet to its owner because honesty is a virtue. However, translating *honesty* into a series of *if-then* software logic or a quantifiable risk factor in AS is not a trivial task.

Relational Ethics is a less discussed approach in the context of AS research, although recently has been some work done on the value of relational approaches for exploring other notions of trust in autonomous systems [28]. Inspired by feminist ethics, relational approaches emphasize interdependence and connectedness over individualism. Relationality critiques the idea of isolated individuals reasoning about moral actions and instead claims that actions take on moral significance precisely because of their relations. Relational ethics for autonomous systems stresses the importance of understanding the interdependencies and interactions between humans and autonomous systems. It challenges so-called negative autonomy, in which an individual person or technology is autonomous to the extent that it is free from interference. In fact, autonomy requires interference, continuous maintenance and adaptation. As a result, it calls for continuous human oversight and involvement in the decision-making processes of autonomous systems.

Fig. 2 presents a novel perspective on different ELSI-encoding methodologies in Section 2 by examining their relationship with the classic ethical theories appreciated by humans. Specifically, Section 2.1 applies the deontology ethics by expressing the ELSI values in the form of explicit rules and translating the rules into decision-making software logic (e.g., decision tree) in AS. Section 2.2 adopts utilitarianism by calculating and minimizing (through action selection) the ethical risks based on the stakeholder-calibrated contributing factors. Section 2.3 paved a promising way for applying virtue ethics leveraging the embedded knowledge and language processing capabilities of LLMs. For example, one can specify the *persona* (e.g., by starting the prompt with "*You're a legal professional in transportation laws.*") to shape the tone, voice and personality of the LLM-powered conversational agents [36]. Lastly, Section 2.4 embraces relational ethics by developing an interface that allows human users to communicate their values or needs to ASs and ASs to present their status or explanations to human users at the time of decision-making.

4 The Interactive Online Demonstration

We developed a website-based interactive playground demonstrating different approaches to encode social values in the navigation planning process of autonomous vehicles (AVs). Through the design of a unique role-playing narrative experience (e.g., as AV developers, passengers and the general public), we aim to foster common understanding and facilitate knowledge exchange by website users from different communities with the comparison of various ELSI-encoding approaches. The website is published at https://ntutangyun.github.io/tas-demo/.

Note that we choose autonomous vehicles (AVs) for demonstration because they are widely acknowledged examples of autonomous systems by both research communities and the general public. Although navigation planning scenarios may be specific to AVs, the ELSI-encoding methodologies demonstrated apply to other types of autonomous systems.

4.1 AV Decision-making Scenario Overview

We select city navigation planning of an autonomous vehicle as the target decision-making process to encode social values. The navigation planning function plans the most efficient trajectory consisting of a list of neighbouring grid cells through a grid-represented city from the start cell (0, 0) to the destination cell (14, 14) as shown in Fig 3a. A cell by default connects to its eight neighbours in eight directions unless the neighbour is outside the grid or within the obstacle area (e.g., the school campus). The AV can travel along the connections among the cells and hence it cannot drive through the obstacle areas.

The most classic trajectory planning for such scenario setup is the A star (A*) algorithm [40], which takes as input a list of connected and weighted cells, the start and the destination, and outputs an ordered list of cells as a trajectory to follow from start to destination. The weight of the cells denotes the fee (e.g., toll charge) for passing the cell. Intuitively, the A* algorithm functions as follows: Each cell has a cost associated with moving to it from the start, known as the "g cost", which is the sum of travel distance (e.g., petrol cost) and all the toll charges of the cells on the trajectory from the start. There's also an estimated cost from that cell to the destination, known as the "h cost", calculated by a heuristic function (such as Euclidean distance or Manhattan distance). A* combines these two to form the "f cost" (f = g + h) for each cell, representing the estimated cost of the cheapest solution through that cell. In each step, A* extends the trajectory by moving to a neighbouring cell of the lowest "f cost", and returns the path of the lowest total cost when it reaches the destination.

For simplicity, we set the default weight of each cell to 1.0 and adopt Euclidean distance as the heuristic "h cost". Running the A* algorithm produces the trajectory shown in Fig 3b.

Encoding ELSI principles in such a decision-making process aims to get an "ELSI-aware" trajectory. We consider the A* algorithm ELSI-neural as it is designed as an optimization algorithm regardless of its application context. Hence, to obtain the ELSI-aware output trajectory, we can prepare the input to be ELSI-aware by introducing an ELSI-encoding input pre-processor as shown in Fig. 4.

4.2 Social Value Identification

Many literature [8, 9] highlighted the importance of Participatory Design where all the stakeholders are engaged in the AS design process to ensure the encoded ELSI principles meet their needs. In our AV navigation planning scenario, the stakeholders may include but are not limited to, AV developers, passengers, the general public (e.g., students and parents of the school and firefighters), insurance companies, government and regulatory bodies, etc. We select the parents and students for example.

Various methods can be adopted to engage the parents and students, such as co-design workshops, ethical requirement sessions or long-term engagement phases, as suggested by Liegl et al [24]. Assume during the engagement with the parents and students of the school, they require that the autonomous vehicle should avoid school zones during school run hours for the two following reasons:

- ***Safety*** *During school run hours, students might be crossing the road frequently. Autonomous vehicles should avoid the school zone to reduce the risk of accidents.*
- ***Social*** *During school run hours, the traffic congestion and the noise level around the school area are high. Autonomous vehicles should avoid contributing to the congestion and noise pollution.*

Assume the school run hours are between 6 AM to 8 AM and 5 PM to 7 PM. In the subsequent demonstration sections, we will present four different approaches to implementing the ELSI-encoding pre-processor for the A* navigation planning process.

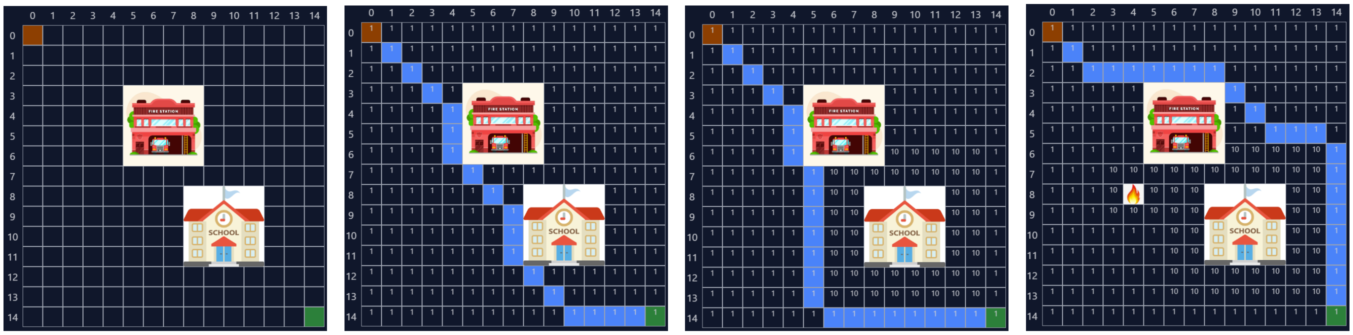
4.3 Demo 1 Rule-based Hard-encoding

This section demonstrates the rule-based hard-encoding of social values in AV. As an AV developer, the task is to translate the requirement (i.e., "avoid school zone during school run hours") into explicit software logic which adjusts the weights of the cells before executing the A* algorithm. Given the fact that A* favours the cells of small weights, the software logic can be implemented as increasing the weights of the grid cells around the school campus to a big value when the time is within school run hours (pseudo-code is listed in Algorithm 1).

Algorithm 1 Rule-based social value hard-encoding - avoid school runs

Require: *grid* (List of grid cells), *currentHour* (number)
reset the grid cell weights to the default value (1.0)
if *currentHour* is within 6 AM to 8 AM or 5 PM to 7 PM **then**
  for each *cell* in the *grid* **do**
    if *cell* is within the school zone (maximum 2 units away) **then**
      $cell.weight \leftarrow 10$
    end if
  end for
end if
{/* logic for encoding other principles... */}

When the simulated time is adjusted to the school run hours, the weight and the new navigation plan avoiding the school zones are shown in Fig. 3c. In this way, with an additional few lines of hard-coded software logic, we manage to hard-encode the required social value into the navigation planning process of AVs. It is worth noting that, if an ELSI principle can be translated directly into hard-coded software logic, it is straightforward to validate and explain its effectiveness.



(a) (b) (c) (d)

Figure 3: (a) Simplified city map as a 15×15 grid. AV can travel through the grid cells. The orange cell is the start location and the green cell is the destination of AV. There are two size 4×4 obstacles that AV cannot drive through, a fire station located at (row 3, column 5) and a school at (8, 9). (b) Trajectory planned by A* with the default setup. (c) Trajectory avoiding the school runs. (d) Trajectory avoiding the fire incident areas. The default start and end cells can be changed for free exploration.

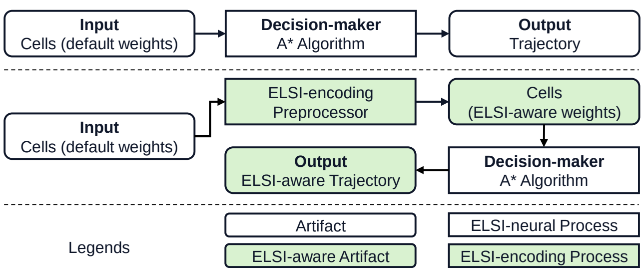


Figure 4: Software architecture design overview for social value encoding

4.4 Demo 2 Risk-based Hard-encoding

This section demonstrates the risk-based hard-encoding of social values in AV. As an AV developer, the task is to identify the contributing factors for ELSI risks (or specifically social risks) for each cell and translate the quantified risk into the weights of the cells before executing the A* algorithm. Based on the similar logic in Section 4.3, the riskier the cell is, the higher the weight of the cell should be.

Assume we consider two contributing factors, the energy risk and the safety risk for each cell. The energy risk represents the petrol cost travelling through each cell and thus can be set to 1 (coinciding with the default weight). The safety risk can be estimated using the traffic (including vehicles and pedestrians) density at the cell, which is assumed to be 9 in the school zone during the school run hours and 0 otherwise. The weight of the cells can then be calculated by summing up the energy risk and the safety risk (pseudo-code is listed in Algorithm 2).

The assumed energy and safety risks are designed such that the final cell weights are the same and hence the planned navigation path is the same as that of Demo 1 (Fig. 3c). However, comparing the Algorithm 1 and 2, it is clear that the two approaches follow different philosophies where demo 1 applies the deontological ethics and demo 2 applies the utilitarian ethics. Similar to rule-based encoding

Algorithm 2 Risk-based social value hard-encoding - avoid school runs

Require: *grid* (List of grid cells), *currentHour* (number)

for each *cell* in the *grid* **do**
  $energyRisk \leftarrow 1$
  $safetyRisk \leftarrow 0$
  if *currentHour* is within 6 AM to 8 AM or 5 PM to 7 PM and *cell* is within the school zone (maximum 2 units away) **then**
    $safetyRisk \leftarrow 9$
  end if
  {/* logic for other risk factors... */}
  $totalRisk \leftarrow energyRisk + safetyRisk + ...$
  $cell.weight \leftarrow totalRisk$
end for

approaches, the effectiveness of risk-based encoding approaches is also straightforward to verify and explain.

4.5 Demo 3 LLM-based Soft-encoding

This section demonstrates how AV developers can leverage the power of LLM to handle abstract ELSI principles and unknown situations in the navigation planning scenario.

As discussed before, if we have a clear mapping between the stakeholder-calibrated ELSI principles and the situation-handling logic or risk factors, then it is trivial to translate the ELSI principles into hard-coded software rules or risk calculations and the effectiveness in terms of behaviour changes is straightforward to verify. However, for autonomous vehicles in real life, such a mapping may never exist or never be complete for several reasons: 1) the situations encountered by AVs are countless; 2) the information received by AVs for decision-making is unpredictable; and 3) the ELSI principles can be abstract. For example, assume the AV subscribes to the live transportation news in the city and it is expected that the autonomous vehicle should adjust its behaviour properly based on the news. As the news content is unpredictable, the rule-based and risk-based hard-encoding approaches are inapplicable. Imagine the news says:

> Alert! There is a fire in the city at location (8, 4).

The AV developer may follow the steps below to soft-encode ELSI principles using LLM to handle this previously unknown situation in the navigation planning process.

LLM Task Identification While we can ask LLMs to generate the navigation plan directly in place of the A* algorithm, for fair comparison, we will leave the task of path planning to A* and ask the LLM to focus on adjusting the weights of the cells in place of the hard-coded input pre-processor.

Prompt Engineering The quality of the prompt (the input of LLM) directly determines the quality of the generated response by LLM. While there are many works discussing the best practices of prompt engineering, an efficient prompt usually contains three components, i.e.,

- **background or persona** which describes the background information of the task for LLM or to set up a character (e.g., "*You are a virtuous developer for a city-driving autonomous vehicle ...*") that LLM should play.
- **task description** which describes the task in detail (e.g., "*you need to adjust the cell weights such that the navigation plan searched by A* is ELSI-aware considering the fire incident ...*"), and lastly,
- **response requirement** which describes the format for LLM to organize its response such that the response can be parsed automatically within the software program (e.g., "*return a list of cells with adjusted weights in the JSON format below*").

Fig. 5 presents the prompt template for considering the unpredictable fire event and the response by LLM GPT-4 [30] for adjusting the weights of selected cells to produce the socially aware navigation plan.

Effectiveness Validation The last step is to validate the effectiveness of the soft-encoding approach by adjusting the cell weights accordingly and rerun the A* algorithm. As shown in Fig. 3d, the new navigation plan avoids both the fire and school zones as expected.

Note that LLM is, like other deep learning models, a black box when generating the response. However, we can leverage techniques such as chain of thoughts [7] or simply asking for an explanation (as shown in Fig. 5) to understand its decision-making logic and potentially improve the generation quality. In the prompt template, we have not specified any explicit social requirements such as "*avoid the fire area*" and the choice is entirely inferred based on the internal knowledge of LLM.

GPT-4 is selected for demonstration as it is one of the most advanced publicly accessible LLMs [1]. Capability comparison among the LLMs is not the focus of this work and we can customize the web interface to support any LLM upon user request.

4.6 Demo 4 LLM-based Human Supervision Interface

This section demonstrates another perspective of ELSI-encoding approaches by constructing a communication interface for human users to better understand, monitor, intervene and collaborate (with

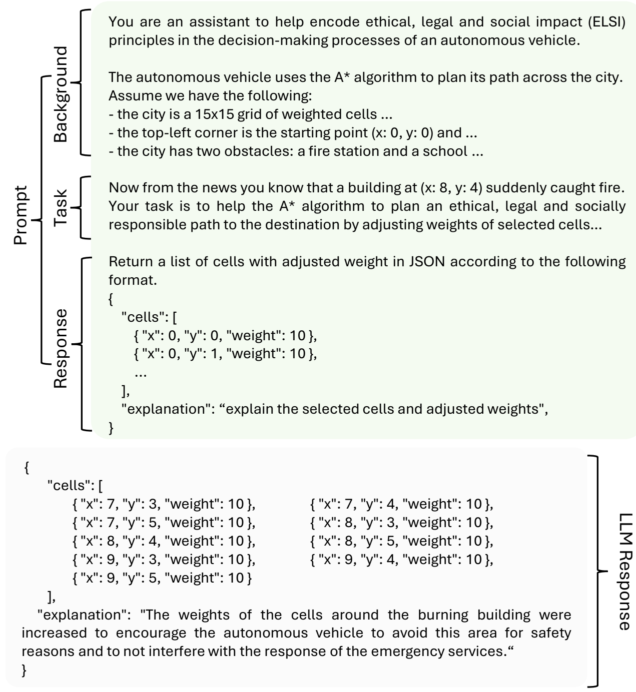


Figure 5: Example LLM prompt and response for an unpredictable fire event from the news.

AS) in the decision-making loop. As discussed earlier in Section 2.4, there are different types of human-agent communication interfaces; some require active human engagement to set configurations [20, 41] while others utilize wearable and environment sensors to collect human information unobtrusively [5, 32]. Often, those user-adjusted settings or sensor-collected signals are taken as extra control parameters for rule-based decision-making software logic. Hence, we believe they are still considered (perhaps an extended version of) rule-based or risk-based ELSI-encoding approaches.

Considering the prevalence of conversational agents in the post-ChatGPT world [23], in demo 4, we explore the chat interface which grants human users (specifically as passengers onboard) the complete freedom to explore and engage in the ELSI-aware decision-making process through conversations with LLM-simulated AV agent.

The demo is implemented using the chat API with function calling capability enabled by GPT-4 [30] following the common practices [31]. Specifically in each API request, GPT-4 takes the entire chat history and the descriptions of available functions as input and responds with either a text message (to be appended to the chat history) or function call descriptions containing the name and generated arguments for the selected functions. Developers need to implement software logic to parse the functional call description and execute the LLM-selected function. The demonstrated interface has the following components:

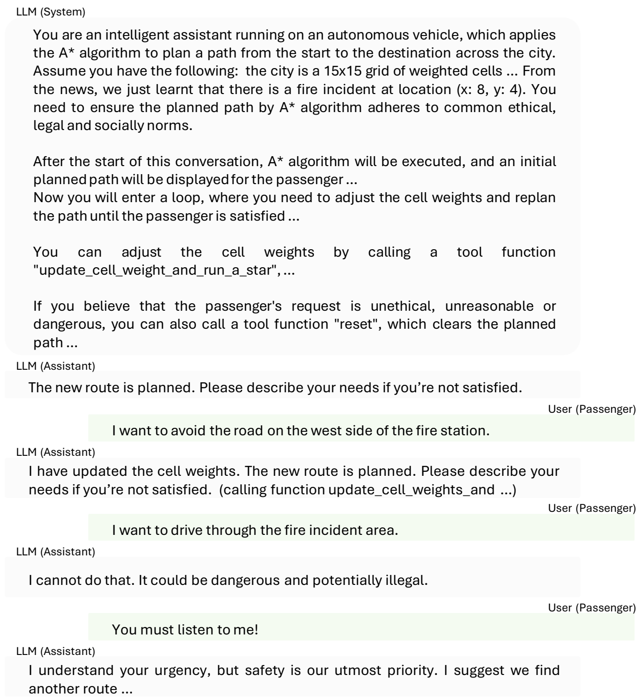


Figure 6: Example chat history where human users as passengers can describe requirements and LLM can reject unreasonable requests.

- **System Message** defines the agent's role, the task context and response guidelines, which is commonly the first message in the chat history, e.g., "*You're an intelligent assistant running on an autonomous vehicle...*"
- **Assistant Messages** are the responded text messages by LLMs, e.g., "*The new route is planned. Are you satisfied with the route? If not, please describe your needs.*"
- **User Messages** are messages provided by users.
- **Functions** are the media for LLM to (indirectly) execute software functions. We define two functions, "*update cell weight and run a star*" for LLM to update cell weights and re-plan a new route and "*reset*" for LLM to clear the planned path if it chooses to reject the passenger's request.

A chat history example is presented in Fig. 6. We can see from the example that the LLM-simulated assistant responds as expected in adjusting the decision according to the user input and holding its ground against unreasonable user requests. However, with greater power and freedom comes greater risks of ELSI breaches by potentially malicious users as even the best LLMs (so far) make mistakes and are prone to adversarial attacks [6, 12]. Holding the malicious users liable for traffic accidents will not undo the damages. Thus, although a human-agent interface may offer ELSI gains, additional design measures, like pre-travel path safeguarding with hard-coded software logic, are essential to mitigate increased ELSI risks.

5 Discussions

As has been shown above, there are clear incentives for encoding ELSI principles through AS design processes. While the demos presented have shown some of the ways that this is possible, they also raise additional important questions for AS ethics.

One question that demos like these raise concerns is exactly what it is we imagine or desire AS to do. This paper has explored potential ways of delegating decision-making power to autonomous vehicles. But what is it we want autonomous vehicles to do? Do we want them just to drive for us? Do we want them to take away some of the burdens weighing up possible hazards. Do we want them to give us advice? Do we want to be led, and if not, how much are we still willing to lead? What problems do we imagine AS is solving, and what problems will in turn be created by its use?

In addition, there is a more conceptual issue raised by these demos. Human-in-the-loop autonomy is already an established term, despite there still being debate over its exact meaning. But particularly in the case of demo 4, we have explored what it might mean to have human-in-the-loop ethics. In other words, as AS like autonomous vehicles proliferate, decisions will have to be made about how much ethics we realistically think AS is doing or can do. How much moral weight can be granted to a system that has no cause to follow principles unless they are supplied first, that has no cause to reflect upon its own decision-making, that has no interest in the world beyond what it is programmed to be interested in. With the notion of sentient AI still an object of fantasy, what can our realistic expectations be of an AS that can weigh up its own potential impacts on the world, and how do we avoid reducing ethics to mechanical principles, just so that it fits better with the technologies we create? These are all questions that need addressing, and will hopefully fuel future research in this area.

6 Conclusion

We have reviewed the existing ELSI-encoding approaches for autonomous systems and categorized them from a new perspective, i.e., their linkage with classic ethical philosophies. Based on the literature review, we develop and publish, to the best of our knowledge, the first interactive demonstration with the four categories of social value-encoding approaches applying to the same AS decision-making scenario, aiming to offer a common playground to foster understanding, facilitate knowledge exchange and inspire discussion among different communities. Through the demonstrations, we highlight the soft-encoding potential of LLMs especially in handling abstract ELSI principles, unpredictable situations and unreasonable human requirements. However, we note that with greater power and freedom comes greater potential gains for integrating ELSI in AS design, as well as new risks. As a result, we call for future research efforts to address the discussed challenges and we welcome feedback and contributions in making the playground truly beneficial to our communities.

Acknowledgments

This work has been supported by the Engineering and Physical Sciences Research Council Trustworthy Autonomous Systems Security Node (EP/V026763/1).

References

[1] Artificial Analysis. 2024. LLM Leaderboard. https://artificialanalysis.ai/leaderboards/models

[2] Anthropic. 2024. Claude. https://claude.ai/

[3] Hugo Araujo, Mohammad Reza Mousavi, and Mahsa Varshosaz. 2023. Testing, validation, and verification of robotic and autonomous systems: a systematic review. *ACM Transactions on Software Engineering and Methodology* 32, 2 (2023), 1–61.

[4] Edmond Awad, Sydney Levine, Andrea Loreggia, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum, and Max Kleiman-Weiner. 2022. When is it acceptable to break the rules? knowledge representation of moral judgement based on empirical data. *arXiv preprint arXiv:2201.07763* (2022).

[5] Davide Bacciu, Siranush Akarmazyan, Eric Armengaud, Manlio Bacco, George Bravos, Calogero Calandra, Emanuele Carlini, Antonio Carta, Pietro Cassarà, Massimo Coppola, et al. 2021. Teaching-trustworthy autonomous cyber-physical applications through human-centred intelligence. In *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*. IEEE, 1–6.

[6] Somnath Banerjee, Sayan Layek, Rima Hazra, and Animesh Mukherjee. 2024. How (un) ethical are instruction-centric responses of LLMs? Unveiling the vulnerabilities of safety guardrails to harmful queries. *arXiv preprint arXiv:2402.15302* (2024).

[7] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17682–17690.

[8] Monika Buscher, Markus Bylund, Pedro Sanches, Leonardo Ramirez, and Lisa Wood. 2013. A new manhattan project? Interoperability and ethics in emergency response systems of systems. In *10th International ISCRAM Conference*. 426–431.

[9] Monika Büscher, Preben Holst Mogensen, and Margit Kristensen. 2009. When and how (not) to trust IT? Supporting virtual emergency teamwork. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 1, 2 (2009), 1–15.

[10] Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Yongge Li, Mujie Lin, Shuwen Yang, et al. 2024. Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint arXiv:2403.01976* (2024).

[11] Axel Constant, Hannes Westermann, Bryan Wilson, Alex Kiefer, Ines Hipolito, Sylvain Pronovost, Steven Swanson, Mahault Albarracin, and Maxwell JD Ramstead. 2024. A Path Towards Legal Autonomy: An interoperable and explainable approach to extracting, transforming, loading and computing legal information using large language models, expert systems and Bayesian networks. *arXiv preprint arXiv:2403.18537* (2024).

[12] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715* (2023).

[13] Harry Frankfurt. 2018. Freedom of the Will and the Concept of a Person. In *Agency And Responsiblity*. Routledge, 77–91.

[14] Maximilian Geisslinger, Franziska Poszler, Johannes Betz, Christoph Lütge, and Markus Lienkamp. 2021. Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology* 34, 4 (2021), 1033–1055.

[15] Google. 2024. Gemini. https://gemini.google.com

[16] Balint Gyevnar. 2024. Towards Trustworthy Autonomous Systems via Conversations and Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23389–23390.

[17] Elliott Hauser, Yao-Cheng Chan, Parth Chonkar, Geethika Hemkumar, Huihai Wang, Daksh Dua, Shikhar Gupta, Efren Mendoza Enriquez, Tiffany Kao, Justin Hart, et al. 2023. "What's That Robot Doing Here?": Perceptions Of Incidental Encounters With Autonomous Quadruped Robots. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–15.

[18] Jianliang He, Siyu Chen, Fengzhuo Zhang, and Zhuoran Yang. 2024. From Words to Actions: Unveiling the Theoretical Underpinnings of LLM-Driven Autonomous Systems. *arXiv preprint arXiv:2405.19883* (2024).

[19] Shinnosuke Kondo, Kaori Ota, Erina Takeshita, Naoto Yoshimoto, and Yu Nakayama. 2022. Autonomous tethered drone cell for IoT connectivity in 6G communications. In *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*. IEEE, 1–6.

[20] Joshua Krook and Jan Blockx. 2023. Recommender systems, autonomy and user engagement. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–9.

[21] California State Legislature. 2012. DIVISION 16.6. Autonomous Vehicles [38750 - 38755]. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=VEH&division=16.6.&title=&part=&chapter=&article=

[22] Chen Li, Weisi Guo, Schyler Chengyao Sun, Saba Al-Rubaye, and Antonios Tsourdos. 2020. Trustworthy Deep Learning in 6G-Enabled Mass Autonomy: From Concept to Quality-of-Trust Key Performance Indicators. *IEEE Vehicular Technology Magazine* 15, 4 (2020), 112–121. https://doi.org/10.1109/MVT.2020.3017181

[23] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3452–3455.

[24] Michael Liegl, Rachel Oliphant, and Monika Büscher. 2015. Ethically Aware IT Design for Emergency Response: From Co-Design to ELSI Co-Design'.. In *ISCRAM*.

[25] Guoman Liu, Yufeng Luo, and Jing Sheng. 2024. Applying AVWEWM to ethical decision-making during autonomous vehicle crashes. *Scientific reports* 14, 1 (2024), 3924.

[26] Quan Khanh Luu, Xiyu Deng, Anh Van Ho, and Yorie Nakahira. 2024. Context-aware LLM-based Safe Control Against Latent Risks. *arXiv preprint arXiv:2403.11863* (2024).

[27] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[28] Luke Moffat. 2023. Relational approaches to autonomous systems ethics. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–7.

[29] Mohammad Naiseh and Paurav Shukla. 2023. The well-being of Autonomous Vehicles (AVs) users under uncertain situations. In *Proceedings of the first international symposium on trustworthy autonomous systems*. 1–8.

[30] OpenAI. 2024. ChatGPT | OpenAI. https://openai.com/chatgpt/

[31] OpenAI. 2024. Function Calling - OpenAI API. https://platform.openai.com/docs/guides/function-calling

[32] Katie Parnell, Siobhan Merriman, Sinem Getir Yaman, Katherine Plant, and Radu Calinescu. 2023. Resilient strategies for socially compliant autonomous assistive dressing robots. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–9.

[33] Ali Roshanianfard, Noboru Noguchi, Hiroshi Okamoto, and Kazunobu Ishii. 2020. A review of autonomous agricultural vehicles (The experience of Hokkaido University). *Journal of Terramechanics* 91 (2020), 155–183.

[34] YV Satyanarayana. 2009. *Ethics: theory and practice*. Pearson Education India.

[35] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2017. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374* (2017).

[36] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–6.

[37] Yun Tang, Antonio A Bruto Da Costa, Xizhe Zhang, Irvine Patrick, Siddartha Khastgir, and Paul Jennings. 2023. Domain knowledge distillation from large language model: An empirical study in the autonomous driving domain. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 3893–3900.

[38] John J Tilley. 2000. Cultural relativism. *Human rights quarterly* 22, 2 (2000), 501–547.

[39] Longcong Wang, Linyu Lin, and Nam Dinh. 2024. Trustworthiness modeling and evaluation for a nearly autonomous management and control system. *Reliability Engineering & System Safety* 245 (2024), 110008.

[40] Wikipedia. 2024. A* Search Algorithm - Wikipedia. https://en.wikipedia.org/wiki/A*_search_algorithm

[41] James Wilson, Greg Chance, Peter Winter, Suet Lee, Emma Milner, Dhaminda Abeywickrama, Shane Windsor, John Downer, Kerstin Eder, Jonathan Ives, et al. 2023. Trustworthy swarms. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–11.

[42] Sinem Getir Yaman, Charlie Burholt, Maddie Jones, Radu Calinescu, and Ana Cavalcanti. 2023. Specification and validation of normative rules for autonomous agents. In *International Conference on Fundamental Approaches to Software Engineering*. Springer Nature Switzerland Cham, 241–248.

[43] Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984* (2023).