EETAS: A Process for Examining Ethical Trade-Offs in Autonomous Systems

Catherine Menon[1][a], Silvio Carta[2][b] and Frank Foerster[1][c]
[1]*Department of Computer Science, University of Hertfordshire, College Lane, Hatfield, U.K.*
[2]*School of Creative Arts, University of Hertfordshire, College Lane, Hatfield, U.K.*
{c.menon, s.carta, f.foerster}@herts.ac.uk

Keywords: Ethics, Design, Ethical Trade-Offs.

Abstract: Public-facing autonomous systems present society with significant ethical challenges, not least of which is the need for stakeholder understanding and discussion of how these systems balance competing ethical principles. In this paper we present EETAS: a structured, gamified process for obtaining stakeholder input into the ethical balances and trade-offs which they consider it acceptable for a proposed autonomous system to make. We describe how outcomes from the EETAS process can be used to inform the design of specified autonomous systems, as well as how the process itself can improve stakeholder engagement and public understanding of ethics in AI and autonomous systems. In support of this we present the findings from an initial EETAS pilot study workshop, which shows an indicative trend of improvement in public understanding and engagement with AI following participation.

1 INTRODUCTION

One of the most complex obstacles to public acceptance of autonomous systems (AS) and AI is the understanding of how competing ethical requirements in these systems may be managed. Standards such as BSI 8611 (British Standards Institute, 2016) on the ethical design and applications of AS, the IEEE guidance on ethically-aligned design (IEEE, 2018) and the Turing Institute guidance on understanding artificial intelligence ethics and safety (Leslie, 2019) provide information to developers on ethical imperatives to be satisfied by the system, but there is very little existing guidance for either stakeholders or developers on managing and understanding ethical complexities and balances.

Furthermore, conversation around AI has traditionally focused on the technology and its capabilities, rather than the diverse ethical concerns of stakeholders and wider society. Nonetheless, autonomous systems cannot exist in an ethical vacuum; rather, they will be expected to conform to the social, legal and ethical norms of the community in which they operate (IEEE, 2018). This is a non-trivial task, as different societies – and different stakeholders within those societies – may prioritise ethical principles differently. For example, societies with a relatively greater regard for governmental authority may be comfortable with autonomous systems which prioritise public safety over data privacy, as demonstrated in the adoption of the TraceTogether app in Singapore (Lee, 2020). Similarly, stakeholders within other societies which prioritise individual choice over public cohesion may have an ethical preference for AS which obey user commands even where this could compromise public safety, such as allowing customisation, or "individuation" into autonomous vehicle technology (Hancock, 2019). In addition to this, societies are of course not homogenous, and individual stakeholders within any given society may also prioritise ethical principles differently depending on age, class, gender and perceived technical competence (Park, 2021).

In this paper we present a structured process for obtaining stakeholder input into the acceptability of ethical trade-offs in a proposed autonomous system. This structured process enables stakeholders to identify potential trade-offs between different ethical

[a] https://orcid.org/0000-0003-2072-5845
[b] https://orcid.org/0000-0002-7586-3121
[c] https://orcid.org/0000-0002-9263-3897

principles and to work collaboratively to identify constraints, or limits, within which they consider these trade-offs to be ethically acceptable. We hypothesise that the benefits of this structured process are two-fold: firstly that participation will improve public understanding of, and willingness to engage with, ethical complexities of AS and secondly that AS designers will gain insight into potential design choices which may be made to render the autonomous system more acceptable to the public.

We also present an interactive tool (Figure 1) which we have created to provide a visual representation of the outcomes of the EETAS process. This tool serves as a record of the public discussion, which can be retained by end-users or stakeholder organisations and used to illuminate diverse public perspectives on AS ethics. In addition, the tool can be used later in the lifecyle to communicate the autonomous system's ethical prioritisations and to increase end users' understanding of it.

In Section 2 we present a discussion of existing literature which considers questions of ethical prioritisation in autonomous systems. Section 3 contains our description of the EETAS process, while Section 4 provides a description of an initial pilot study workshop which has demonstrated an indicative trend between participation in EETAS and enhanced public understanding of AS. Section 5 identifies our conclusions and some steps for further work.

2 BACKGROUND

The concept of trade-offs, or risk balancing, between two desirable properties is well-established as a research area. Expected utility theory (von Neumann, 1947) describes how an individual's general attitude to risk and benefits can change their willingness to accept particular specified risks. Similarly, prospect theory (Kahneman & Tversky, 1979) also allows a more complex framing of risk perception and risk appetite. The trolley problem (Foot, 1967) is of course perhaps the seminal example of risk balancing, and has informed much of the public discourse around autonomous vehicle behaviour.

Beyond this, risk balancing as a concept is well-explored in autonomous system development. (IET, 2019) describes trade-offs between safety and security of cyber-physical systems, while (Akinsanmi, 2021) considers the balancing of public health, privacy and digital security. Within specific autonomous domains the concept of prioritising certain safety or ethical properties has also been discussed: (Thornton, 2018) describes the tension between the desire for personal autonomy on the part of an autonomous vehicle user, and the more general desire for fairness and public safety while (Lin, 2015) also considers how specific actions on the part of an autonomous vehicle – e.g. driving closer to another car in order to give more room to a pedestrian – transfer the risk from one segment of the population (pedestrians) to another (other drivers). In the field of healthcare, ethical trade-offs between privacy and well-being are also common (Lee, 2020), (Martinez-Martin, 2020).

Other existing work focuses specifically on trade-offs which affect the design process. (Dobrica, 2002) presents a comprehensive survey of trade-offs in complex systems design, while (Goodrich, 2000) discusses these trade-offs within an autonomous context, specifically that of collision avoidance systems. Similarly (Bate, 2008) considers trade-offs more generally within safety-critical systems, while (Menon, 2019) proposes a methodology for developers of autonomous vehicles to justify and communicate the ways in which their system design has been informed by ethical trade-offs.

The benefit of using a tangible element such as the interactive tool in Figure 1 to test and visualise trade-offs in real-time is supported by a large body of literature, including (Schrier, 2019), (Rossi, 2019), (Larson 2020). More generally, games have been shown to be a successful vehicle for engagement with ethics principles, especially in industry testing. Examples include Judgment Call, (Ballard, 2019), a game developed to help AI developers to identify ethical questions using design fiction, as well as MiniCode, a design fiction toolkit developed for near-future technology designers and developers (Malizia, 2022).

Much of the existing work around autonomous systems and AI is focused on developers, intended either to provide them with insight into how a system can be designed or to be used as guidance on making ethically justifiable decisions. However, there is comparatively little work which provides stakeholders and end-users with an opportunity to express their concerns around AI ethics, or to inform the design of a proposed system by providing input into the perceived acceptability of ethical trade-offs. The process we describe here addresses this gap.

3 EETAS PROCESS

In this section we describe a process for obtaining stakeholder input into decisions about the ethical prioritisations embedded into autonomous systems. It is important that the EETAS process takes place relatively early in the design of such systems, to allow the outcomes to be fed back into the design lifecycle and consequently enable developers to integrate ethically acceptable behaviour into the system from the ground up. In consequence, the system may not be fully specified at the time the EETAS process takes place. This is expect, and the process allows for under-specification and early prototypes of an autonomous system to be used.

3.1 Step 1: Provide AS Description

A participant group is selected, including developers of the AS under consideration, proposed end-users, regulators and members of the public. The developers provide the group with a written, accessible description of the AS and its relevant functions. Appropriate descriptions may specify, for example, that this is "an assistive robot that reminds you when to take medication, alerts you when you have left the oven on and engages you in conversation". It is likely that participants will have further questions around the functionality of the system – e.g. "does the assistive robot speak to me or do I access it via a screen?" – and these should be clarified with the developers as part of this step. As EETAS is ideally undertaken during the early stages of development (e.g. requirements gathering or design), it is likely that some questions around specific functionality cannot yet be answered. These, in turn, become the seeds for the scenarios that teams will identify in Step 3.

3.2 Step 2: Identify Relevant Ethical Principles

The participants are divided into teams of between 4 and 10. This follows (Curral, 2001), with the intent of ensuring diversity of perspectives while still enabling effective and equable dialogue. Each team is provided with a set of pre-prepared cards listing ethical and ethically-informed functional properties which may be desirable for this type of autonomous system. The ethical properties in this set have been identified from a literature review of existing and developing standards, including (BSI, 2016), (IEEE, 2018), (Leslie, 2019), (National Cyber Security Centre, 2019). Some sample principles, which we used in the initial pilot validation workshop (Section 4) are:

- System promotes human physical safety
- System obeys human commands
- System promotes affinity with human user
- System maintains data privacy
- System is accurate
- System is fair
- System maintains human autonomy
- System promotes human long-term health

We note that not all the ethical or functional properties in the full set will be relevant for every system, and that for specific systems there may be additional ethical or functional properties. To address this teams are also provided with a set of blank cards and are encouraged to "tailor" the set of ethical properties to discard those they consider irrelevant, and identify any others considered relevant to this specific system, using techniques such as collaborative discussion, brainstorming and if-then thinking.

3.3 Step 3: Scenario Construction

Teams are then asked to generate scenarios in which two of the set of ethical properties are in conflict with each other during system operation. For example, a team may postulate a scenario where: "the user asks their assistive robot not to remind them about medication today, because they don't want to take it". In this scenario the ethical properties of "system promotes human long-term health" and "system obeys human commands" are in conflict. Similarly, allowing teams to tailor the set of ethical principles may give rise to a scenario for a robot doctor where the system attempts to "engender trust in the human user" by mimicking human appearance and gestures to make the patient feel at ease, thereby causing conflict with another ethical property: "system does not attempt to deceive".

To assist in generating the scenarios, teams are provided with a ready-made checklist of guidewords, to be applied in turn to each of the ethical properties. These guidewords enable teams to work collaboratively to brainstorm scenarios, following the principles of Hazard and Operability Analysis (HAZOP) studies (BSI HAZOP, 2016). The guidewords are presented in Table 1.

Teams should remember that the intent is to identify scenarios in which two or more ethical principles are in conflict: it is not sufficient to identify scenarios which themselves simply represent ethical hazards.

Table 1: HAZOP guidewords for EETAS.

Guide word	Meaning
TOO MUCH	Ethical conflict arising from a scenario where the robot performs its functions in such a way that it grants this ethical property to too many people / in too many circumstances / to too high a degree
NOT ENOUGH	Ethical conflict arising from a scenario where the robot performs its functions in such a way that it grants this ethical property to too few people / in too few circumstances / to too restricted a degree
UNIFORMLY	Ethical conflict arising from a scenario where the robot performs its functions in such a way that the outcome is applied uniformly to everybody / is applied in exactly the same way to everybody
INCONSISTENTLY	Ethical conflict arising from a scenario where the robot performs its functions inconsistently / differently for different people / differently each time
UNEQUALLY	Ethical conflict arising from a scenario where the robot performs its functions such that the beneficial outcome applies only to some people

Participants can be encouraged to apply creativity when identifying scenarios, and may find it helpful to consider the following questions:

- Who is the user of the system?
- Who would be negatively affected in this scenario?
- Who would benefit in this scenario?

3.4 Step 4: Identify Constraints

Teams are then asked to swap scenarios with each other. Each team then works collaboratively to identify design, environmental or end-user constraints under which they would accept different ethical balances in each of the provided scenarios. To assist in this activity, we suggest participants should consider the following questions:

- Which outcome, or balance of outcomes, would you prefer in this scenario?
- Would you accept any alternate outcome in this scenario if users were told beforehand that this is how the system operates? What about if the general public were told beforehand?
- Do you think the trade-off in this scenario is appropriate given the corporate goals and strategy of the design organisation?
- Do you think the person benefitting from different balances of outcomes in this scenario has the moral right to do so?
- Could some of the ethical trade-offs described in this scenario be acceptable in a different environment? With different users? If these did not impact the same people?
- Is there more information which you would need in order to accept some of the possible ethical trade-offs in this scenario?

To gamify this step, each team is allocated points for every scenario in which they identify constraints that render at least two different balances of ethical principles acceptable. Teams should be asked to vote on whether they think these constraints are feasible to implement, and additional points allocated accordingly.

3.5 Use of Design Tool

Steps 3 – 4 are to be performed with the aid of a pre-prepared design tool, EETAS-Trade-Offs-for-You (EETAS-TOY), which represents the AS by a solid block and the relevant ethical principles as sliding bars, as in Figure 1. Participants connect bars end-to-end to represent ethical trade-offs and to discuss how different principles may be prioritized in each scenario.

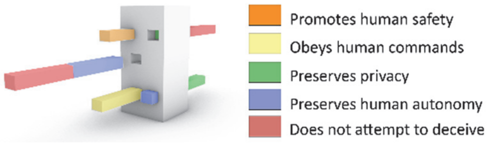

Figure 1: The EETAS-TOY gamified tool.

Bars may be connected to other bars further down the structure to represent where a single ethical property (e.g. "maintains privacy") is implicated in multiple trade-offs (e.g. in balance with both "maintains security" and "explains decisions"). Should participants wish to decouple the two bars representing these two trade-offs, this can be done by identifying a design requirement which permits the decoupling of those aspects of the design which can provide privacy at the cost of security, and privacy at the cost of explainability.

3.6 Recording Outcomes

The outcomes of each step are to be recorded using techniques such as mind-mapping (Beel, 2011). These records can then be used by the AS developers to identify further design requirements which enable or implement the constraints identified by each team. The EETAS-TOY tool itself may be retained by user organizations to aid in explaining ethical trade-offs or to act as a public record of the conversation.

4 PILOT STUDY VALIDATION

We conducted an initial pilot study workshop to investigate participants' perception of the EETAS process and its effect on their own understanding of ethical complexities in autonomous systems. As a preliminary pilot study, this workshop aimed to provide a partial validation of the EETAS process by establishing a link between EETAS participation and understanding of, and willingness to engage with, ethical complexities of AI. A further, orthogonal, aim of the study was to investigate the extent to which the EETAS-TOY tool was perceived as helpful in facilitating discussion and communication amongst participants about the ethical complexities and balances in autonomous systems.

The experiment was approved by the University of Hertfordshire's Health, Science, Engineering and Technology Ethics Committee under protocol number SPECS/SF/UH04940.

4.1 Pilot Study Design and Methodology

The pilot study was carried out at the University of Hertfordshire, with participants recruited following self-selection into the study. After obtaining consent, participants were randomly divided into teams of 4 – 5. The random assignment was performed by the researchers in order to mitigate against the confounding effects of team members knowing each other, or sharing demographic characteristics. All participants were provided with an overview of the purpose of the EETAS process and the workshop, but were not introduced to each individual step of the process in advance, in order to avoid anticipation of some of the discussion points.

All teams were given a high-level written descriptive specification of the robot chosen for consideration throughout the workshop: an assistive robot for use in a domestic environment. Owing to time constraints, a real-world robot prototype and specification could not be sourced for the workshop, and instead the specification was produced by the researchers and based on previous work carried out at the University of Hertfordshire Robot House (Menon, 2019), (Koay, 2020), (Saunders, 2016).

Some functionality ascribed to the assistive robot within this written specification included:

- Moving about the house in response to user commands or actions
- Reminding the user to take medication
- Engaging the user in social interaction or conversation
- Notifying the user of hazardous conditions such as the oven being switched on
- Communicating with other smart device sensors in the house, including camera, audio and personal computers
- Communicating warnings to external medical monitoring systems regarding the health and activities of the user

Participants were given a set of pre-printed cards containing the eight ethical principles described in Section 3.2. Owing to time constraints, all teams were instructed to consider only these ethical principles throughout the workshop and not to expand their selction. Each team was also given an EETAS-TOY tool (Figure 1) and shown how this could be used to represent ethical trade-offs and balances.

Participants were provided with an initial questionnaire and asked to provide information on age, gender and whether they had any background relating to either design or robotics. They were also asked to rank the eight ethical principles in order of how important they considered each of them to be for the assistive robot under consideration.

Following this, teams were introduced to each other and took part in a small ice-breaker. They were then instructed to complete Steps 3 and 4 of the EETAS process, Steps 1 and 2 having been completed by the researchers (owing to time restrictions, the HAZOP guidewords were not used). Teams were allocated 20 – 35 minutes for each step, with the researchers indicating when the time for each step was nearing completion. All teams were given structured worksheets to record their identified scenarios (Step 3), and constraints (Step 4, following swapping of team records).

Teams worked simultaneously in different parts of the workshop room, with each team being observed and monitored by one of the researchers. The researchers were able to answer questions and remind participants of the requirements of each step, but did not contribute to the discussions or guide them in any way.

4.2 Post-Study Questionnaires

Following the workshop, participants were asked to complete some further post-study questionnaires. These included the following questions:

- Participants were asked to rank the eight ethical principles in order of how important they now considered each of them to be for the assistive robot
- Participants were asked to give a numerical score of how well they understood ethical trade-offs before the EETAS process, and how well they understood these trade-offs following EETAS (0 = not at all, to 5 = very well)
- Participants were asked to give a numerical score of how helpful they found the EETAS process in understanding ethical trade-offs (0 = unhelpful, to 5 = very helpful)
- Participants were asked to give a numerical score of the EETAS-TOY tool in a) understanding and b) communicating about ethical trade-offs (0 = unhelpful, to 5 = very helpful)

4.3 Pilot Study Results

As this was a preliminary study, with correspondingly low participant numbers (<20), no statistical significance between conditions and questionnaire responses was expected. Nevertheless, there were indicative trends to support our hypothesis of a causal relationship between participation in the EETAS process and improved public understanding of AI ethical complexities.

4.3.1 Participant Demographics

Participant selection was strongly biased towards both design and robotics, with 93% of participants identifying as having a background in design, and 43% a background in robotics. This commonality in background is due to constraints around the recruitment and identification of participants, with most participants sourced via existing connections to the University of Hertfordshire. The age range of participants was 19 – 61 years old, with the average age being 37. The gender balance was roughly equal, with 57% male participants and 43% female.

4.3.2 Participant Responses to EETAS

As may be expected, prior to the workshop participants without a robotics background rated their existing understanding of ethical trade-offs in autonomous systems as lower (mean value 2.6) than those with a robotics background (mean value 3.4). Post-workshop, the gap had narrowed, with those from a non-robotics background rating their understanding of ethical trade-offs as an average of 3.8, compared with 4.2 for those from a robotics background. This corresponds to an increase of 58% greater improvement in understanding ethical trade-offs for those without a robotics background, as compared to those with.

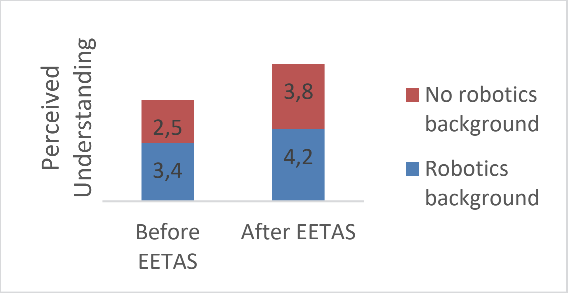

Figure 2: Change in perceived understanding of ethical trade-offs as a result of EETAS participation.

When asked to identify how helpful the process was in understanding ethical trade-offs (0 = unhelpful, to 5 = very helpful), 94% of participants ranked the helpfulness of the EETAS process at 3 or above, with the mean ranking being 3.7. Interestingly, there was no difference noted in this result between those with a robotics background and those without.

When asked about the helpfulness of the EETAS-TOY tool in identifying ethical trade-offs, 64% of participants ranked this as 3 or above (mean value 3.3) and when asked about the helpfulness of the tool in discussing ethical trade-offs, 71% of participants ranked this as 3 or above (mean value 3.7). In contrast to the scores for the perceived helpfulness of EETAS, which were independent of background, those without a robotics background considered the tool more helpful than those with.

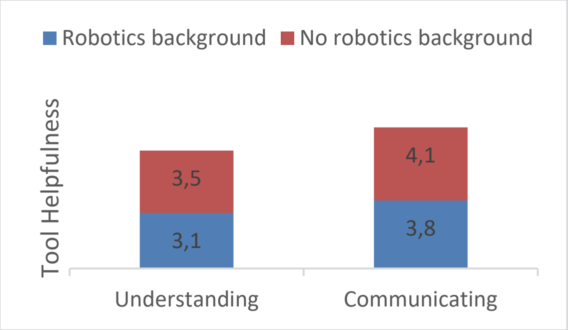

Figure 3: Perceived helpfulness of the EETAS-TOY tool in understanding and communicating.

4.3.3 Observations and Timing

Three teams were each monitored by a researcher, who recorded the total time taken for each step in the EETAS process, and the total time spent using the EETAS-TOY tool. Only time spent actively and purposely using the tool in discussion was recorded, and time spent "fidgeting" with the tool or learning how to use it was discarded.

On average, teams used the EETAS-TOY tool during 45% of the time they were engaged in Step 3 (scenario construction). Teams did not use the tool to any significant extent in Step 4 (identifying constraints). There was no correlation noted between background in either design or robotics and readiness to engage with the tool.

4.3.4 Discussion and Indicative Trends

Although no statistically significant conclusions can be drawn, the results demonstrate some potential indicative trends. Firstly, the EETAS process was considered by a large majority of the participants to be helpful in understanding and discussing ethical trade-offs. This was not correlated with prior experience: those with a robotics background found it to be as helpful as those without. This supports our initial hypothesis that EETAS can be used to improve public understanding of, and engagement with, ethical complexities in autonomous systems.

Moreover, all participants considered that their understanding of ethical trade-offs in autonomous systems had increased following participation in the EETAS process. In this case the extent of the effect could be seen to be correlated with prior experience: those without a robotics background considered that their increase in understanding was greater than those with. This indicates that the EETAS process may serve a useful purpose in raising understanding of autonomous system ethics amongst those who have traditionally been marginalised, or excluded from, existing conversations around AI.

Finally, participants considered the EETAS-TOY design tool to be useful in identifying ethical trade-offs, and discussing these within their teams. Observational monitoring supported an indication that the tool appears to stimulate positive interaction amongst participants by providing a physical aid to visualise trade-offs. We consider it likely that the tangible element of the tool is of value here in supporting participants in abstract reasoning and discussion of unfamiliar concepts.

5 CONCLUSIONS

We have presented a structured, collaborative process for improving public engagement and understanding of ethical complexities in autonomous systems, and for consequently informing the design of these systems. Our results from an initial pilot study workshop indicate that the process improves understanding of ethical trade-offs, most significantly for those who do not have prior experience in robotics or autonomous systems. Our results also indicate that the process is helpful in stimulating debate and discussion around ethical trade-offs, particularly amongst groups of people who have a varied prior understanding of the issues.

We have also considered the use of a physical design tool, the EETAS-TOY tool, in helping participants understand and communicate about ethical trade-offs in a specified system. Our results indicate that participants consider this a helpful physical aid to concretize some of the more abstract concepts, an effect which is more pronounced amongst those without a background in robotics or autonomous systems.

In terms of next steps, we plan to run a larger workshop involving developers and a more varied participant group (e.g. stakeholders, regulators, end-users). We intend to use a real-world robot prototype within this workshop, in order to assess the effectiveness of EETAS in producing outcomes which can be used by the developers to inform the design of the robot.

We also plan to explore the design space of the EETAS-TOY tool more fully. We anticipate developing multiple versions and configurations of the tool, in order to assess the beneficial effects of different tools on participant understanding. New versions will include tools with automated mechanisms to adapt to participant choices, tools with different shapes to assess the importance of visual balance and tools with simplified shapes and interaction methods to aid accessibility.

REFERENCES

Akinsanmi, T., Salami, A. (2021). *Evaluating the trade-off between privacy, public health safety, and digital security in a pandemic*, In Data and Policy 3(27).

Ballard, S., Chappell, K.M. and Kennedy, K. (2019). *Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology*. In Proceedings of the 2019 Designing Interactive Systems Conference, pp. 421-433.

Beel J., Langer, S. (2011). *An Exploratory Analysis of Mind Maps*. In Proceedings of the 11th ACM Symposium on Document Engineering, pp. 81-84.

British Standards Institute. (2016). *Guide to the ethical design and application of robots and robotic systems,* BSI 8611.

British Standards Institute (2016). *Hazard and Operability Studies: Application Guide*, BSI 61882.

Curral L., Forrester, R., Dawson, J., West, M. (2001). *It's What You Do And The Way You Do It: Team Task, Team Size and Innovation-related Group Processes*. In European Journal of Work and Organizational Psychology, 10(2), pp. 187-204.

Foot, P. (1967). *The problem of abortion and the doctrine of double effect*, Oxford Review 5, 5—16.

Hancock, P., Nourbakhsh, I., Steward, J. (2016). *On the future of transportation in an era of automated and autonomous vehicles*. In Proceedings of the National Academy of Sciences, 116(16), pp 7684 – 7691.

Institute of Electrical and Electronic Engineers (2018). *Ethically Aligned Design*, IEEE, v2, https://standards.ieee.org/wpcontent/uploads/import/documents/other/ead_v2.pdf

IET, (2019). *Code of Practice: Cyber Security and Safety*, Available at https://electrical.theiet.org/guidance-codes-of-practice/publications-by-category/cyber-security/code-of-practice-cyber-security-and-safety/

Kahneman, D., & Tversky, A. (1979). *Prospect theory: An analysis of decision under risk*. Econometrica: Journal of the Econometric Society, pp. 263-291.

Koay, K., Syrdal, D., Dautenhahn, K., Walters, M..(2020). *A narrative approach to human-robot interaction prototyping for companion robots*, in Paladyn, Journal of Behavioural Robotics, 11, pp. 66 – 85.

Larson, K. (2020). *Serious games and gamification in the corporate training environment: A literature review*. TechTrends, 64(2), pp.319-328.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*, The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529.

Lee, L., Lee, H. (2020). *Tracing surveillance and auto-regulation in Singapore: 'smart' responses to COVID-19*. In Media International Australia, 177(1), pp. 47-60.

Lin, P. (2015). *Why Ethics Matters for Autonomous Cars*. In Autonomes Fahren, Springer Vieweg, pp 69 –85.

Malizia, A., Carta, S., Turchi, T., Crivellaro, C. (2022). *MiniCoDe Workshops: Minimise Algorithmic Bias in Collaborative Decision Making with Design Fiction.* In Proceedings of the Hybrid Human Artificial Intelligence Conference.

National Cyber Security Centre (2019). *Intelligent Security Tools,* Available at https://www.ncsc.gov.uk/collection /intelligentsecurity-tools

Park, J., Hong, E., Le, H. (2021). *Adopting autonomous vehicles: The moderating effects of demographic variables*. In Journal of Retailing and Consumer Services, 63.

Rossi, F., Mattei, N. (2019). *Building ethically bounded AI. In Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), pp. 9785-9789.

Saunders, J., Syrdal, D., Koay, K., Burke, N., Dautenhahn, K. (2016). *Teach Me - Show Me' - End-user personalisation of a smart home and companion robot*. In IEEE Transactions on Human-Machine Systems, 46(1), pp. 27–40.

Schrier, K., (2019). *Designing games for moral learning and knowledge building*. Games and Culture, 14(4), pp.306-343.

Thornton, S. (2018). *Autonomous vehicle motion planning with ethical considerations*, (Doctoral dissertation), Stanford University.

Von Neumann, J. & Morgernstern, O. (1947). *Theory of Games and Economic Behaviour*. Princeton University Press.