This is a repository copy of *Safe, Ethical and Sustainable : Framing the Argument*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/198044/

Version: Published Version

Proceedings Paper:
McDermid, John Alexander orcid.org/0000-0003-4745-4272, Burton, Simon and Porter, Zoe (2023) Safe, Ethical and Sustainable : Framing the Argument. In: Parsons, Mike, (ed.) The Future of Safe Systems:Proceedings of the 31sr Safety-Critical Systems Symposium, 2023. Safety Critical Systems Club , UK , pp. 297-316.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Safe, Ethical and Sustainable: Framing the Argument

John A McDermid[1], Simon Burton[2], Zoe Porter[1]

[1]Assuring Autonomy International Programme

University of York, UK

[2]Fraunhofer IKS

Munich, Germany

Abstract *The authors have previously articulated the need to think beyond safety to encompass ethical and environmental (sustainability) concerns, and to address these concerns through the medium of argumentation. However, the scope of concerns is very large and there are other challenges such as the need to make trade-offs between incommensurable concerns. The paper outlines an approach to these challenges through suitably framing the argument and illustrates the approach by considering alternative concept designs for an autonomous mobility service.*

1 Introduction and motivation

In many domains, e.g. aviation, there have been long-term improvements in safety. This can be seen as a desirable consequence of consumerism (McDermid et al 2022) amongst other forces. There is now a growing interest in other characteristics of systems such as their environmental and ethical impacts. This is, in part, because safety is often now seen as a given and because the greater scope and capability of modern systems introduces other concerns, such as unfair distribution of risk amongst stakeholders. To reflect this wider set of concerns, we have previously suggested that there is a need to design and assure systems, so that they are 'safe, ethical and sustainable' (McDermid 2022). But this is easier said than done.

© JA McDermid, S Burton, Z Porter 2023.
Published by the Safety-Critical Systems Club. All Rights Reserved.

The aim of this paper is to indicate how we might achieve this goal and to demonstrate that we have done so. There are several challenges here. First, the set of concerns encompassed by 'safe, ethical and sustainable' is vast – water quality, consumption of rare-earth elements/metals, deforestation, ozone depletion to mention but a few of the possible environmental (sustainability) issues. Second, the systems can be very complex, including use of artificial intelligence (AI) so analysing them and their impacts (consequences) is a major undertaking. Third, there may be conflicts between, say, safety and ethical treatment of specific stakeholder groups. There is a trade-off here between utilitarian safety aimed at maximising safety over an entire group (e.g. all road users) against ensuring that each individual has an equal right to safety. This thinking can be extended further, e.g. technologies that improve road traffic safety, but rely on rare earth metals may lead to damage to the environment and exploitation of emerging economies thus harming future generations. Fourth, the set of concerns are often incommensurable in the sense that the units for measuring safety, e.g. risk or rate of occurrence of accidents, are very different from those for, say, air quality which has several measures, e.g. particulate density; similarly, measures for mental and physical harms are also quite different.

The approach to those challenges proposed here is to consider systems at the level, or stage, of their conceptual design and to employ argumentation to illuminate and justify the trade-offs between different system properties – but doing so depends on 'framing the argument' to reduce the potential concerns to a feasible set.

The rest of this paper is structured as follows. Section 2 considers the potential sets of concerns encompassed within 'safe, ethical and sustainable' and takes a first step towards bounding the scope of analysis. Next, section 3 draws on work on governance of complex systems (Burton et al 2021) as a way of 'framing the argument' which is later articulated by drawing on work on ethical assurance arguments (Porter et al 2022). Section 4 uses provision of mobility as a service (MaaS) as an illustration, including addressing different system attributes and stakeholder concerns within the argumentation framework. Conclusions are presented in section 5.

2 Life, the universe and everything

Douglas Adams famously said that '42' was the answer to 'life, the universe and everything' (Adams 1982). In saying that systems should be 'safe, ethical and sustainable' we are perhaps considering a slightly narrower problem – albeit one which doesn't seem to have such a simple answer! As a step towards framing the argument for 'safe, ethical and sustainable' we first consider some very broad articulations of concerns. More specifically, for this audience we take the meaning of safety to be understood and expand more on ethical principles and sustainability before considering a synthesis of the two sets of concerns. Finally, we discuss one area where we need to 'dig deep' – rare earths. As well as being a concern in itself, this illustrates the interdependencies which need to be understood when making trade-offs.

2.1 Ethical principles and artificial intelligence

In practical ethics, it is common to distinguish consequentialist approaches which focus on outcomes, and which take the right actions to be those that bring about the best consequences, and deontological approaches, which focus on duties, and which take right actions to be those that conform to some rules, e.g. do not steal. It is quite hard to relate to such concerns in the context of system design and development. However, as there have been some egregious cases of unethical consequences with systems based on AI and machine learning (ML), in recent years numerous public and private sector organisations have released sets of ethical principles for the development and deployment of AI (Fjeld et al 2020; Jobin, Ienca & Vayena 2019).

In a previous analysis (Porter et al 2022) we have drawn on insights in the ethical AI literature (Floridi and Cowls 2021) to argue that the content of these many sets of ethical principles for AI – which cover values such as fairness, safety, well-being, and privacy, to name but a few - lends itself to a more simplified framework of four ethical principles which have their origin in biomedical ethics (Beauchamp and Childress 1979). These four principles are: beneficence; non-maleficence; respect for personal autonomy; and justice. The four principles can be adapted to the ethical concerns about advanced technologies, when supported by transparency of the assurance argument as well as the ML elements themselves. Our understanding and adaptation of the four principles in the AI context is as follows:

1. beneficence (the system should bring benefit to stakeholders);
2. non-maleficence (the system should avoid unjustifiable harm to stakeholders either directly or indirectly, e.g. via environmental effects);
3. respect for human autonomy (stakeholders should have appropriate and meaningful control over the system or its impact on them);
4. justice (there should be a fair or equitable distribution of benefit and risk from the system across stakeholders).

This underpins the framing of the argument in section 3.

2.2 Sustainable development

The UN General Assembly adopted the 17 Sustainable Development Goals (SDGs) during their 70th session in 2015, as illustrated in Figure 1.

Fig. 1. Sustainable Development Goals

Some of these are essentially individual concerns, e.g. 2 & 3; some are societal, e.g. 8 & 9; many are about the environment and focus more directly on sustainability of the planetary ecosystem. The UN itself prompts international action on these goals, e.g. in October 2022 they focused on goal 2, zero hunger (UN 2022), and other organisations, e.g. UNESCO, link some of their activities to the SDGs.

But if we are designing a particular system, how can we address such global concerns? More narrowly, is it possible to work out which concerns our system might impact? The second question is easier to answer. For example, for an urban autonomous transport system, perhaps the primary SDG is 11 (sustainable cities and communities), and others such as 10 (reduced inequalities) would apply in terms of access to transportation. Some of the SDGs, e.g. 12 (responsible consumption and production) apply to almost all systems. We consider the first question in section 3 under framing.

2.3 The doughnut economy

Perhaps surprisingly, some of the thinking that has tried to blend the perspectives of sustainable development and the (ethical) needs of individuals has come from economists. The ideas are credited to Kate Raworth (Raworth 2017) and are set out in Figure 3, although the simplified 'doughnut' (Wiedmann et al 2020) in Figure 2 is a better starting place for understanding the concepts.

Fig. 2. The Doughnut Economy: Simplified View (Wiedmann et al 2020)

The basic idea is that the acceptable – sustainable, safe and just – space for humanity (including our dependence on the planet) is in the green ring[1]. Outside this ring there are environmentally unsustainable consequences, e.g. through global warming. In the centre there is unacceptable poverty and other unmet needs, e.g. lack of clean water. Taking a radial slice through the figure, this can be seen as analogous to the ALARP (As Low As Reasonably Practicable) concept – but with two ALARP triangles, one where the outside is unacceptable environmentally and the inside is unacceptable from an individual or societal perspective.

Figure 3 shows a more detailed version of the model, with subdivision of the inner circle and outer annulus into more specific concerns. Elements of the inner circle can be seen to correlate with the SDGs e.g. gender equality (SDG 5) and water (SDG 6). There are similar, although less obvious, correlations between the outer annulus and the SDGs, e.g. air pollution relates to climate action (SDG 13). But one of the reasons for showing this more detailed figure is to illustrate the difficulty of identifying and agreeing even the top-level concerns.

[1] This is referred to as 'doughnut economics' although the acceptable, green part, of the Figure is an annulus not a torus.

Fig. 3. The Doughnut Economy (Raworth 2017)

2.4 Rare earths

The above discussion is all high-level – it is time to come down to (rare) earth(s). Affordable and clean energy (SDG 7) depends on renewable energy sources, e.g. solar and wind, both of which are inherently variable. Also, sustainable cities (SDG 11) need sustainable energy and batteries for electric vehicles (EVs) – unless we design '15-minute cities'[2] – and this is perhaps only implicit in the SDGs (e.g. 9, 10 and 11). Such dependence on renewables is even less obvious in the doughnut economy (energy, income and work, and social equity perhaps). So, as well as the high-level structures given by the SDGs and doughnut economics we need to 'dig down' into the supply chain, including considering the rare earths needed for renewables, batteries, etc.

An analysis of what would be required to replace fossil fuels for power generation and for transport (Michaux 2021) identified significant problems with materials including rare earths, see Figure 4[3]. In several cases, the estimate is that it will take millennia at current extraction rates to produce enough materials to replace fossil fuels. Note that this is for just one generation of systems, e.g. EVs or wind farms, and does not consider replacements – where reuse/re-cycling would be vital.

[2] Where all citizens' needs can be met within a 15-minute walk or cycle ride from their homes.

[3] The figure comes from an associated presentation (Michaux 2022), not the report itself.

Metal	Element	Total metal required produce one generation of technology units to phase out fossil fuels (tonnes)	Global Metal Production 2019 (tonnes)	Years to produce metal at 2019 rates of production (years)
Copper	Cu	4 575 523 674	24 200 000	189,1
Nickel	Ni	940 578 114	2 350 142	400,2
Lithium	Li	944 150 293	95 170 *	9920,7
Cobalt	Co	218 396 990	126 019	1733,0
Graphite (natural flake)	C	8 973 640 257	1 156 300 ♦	3287,9
Graphite (synthetic)	C		1 573 000 ♦	-
Silicon (Metallurgical)	Si	49 571 460	8 410 000	5,9
Vanadium	V	681 865 986	96 021 *	7101,2
Rare Earth Metals	-			
Neodymium	Nd	965 183	23 900	40,4
Germanium	Ge	4 163 162	143	29113,0
Lanthanum	La	5 970 738	35 800	166,8
Praseodymium	Pr	235 387	7 500	31,4
Dysprosium	Dy	196 207	1 000	196,2
Terbium	Tb	16 771	280	59,9

Fig. 4. Materials Needs for Replacing Fossil Fuels (Michaux 2022)

Further, there are environmental impacts of mining – for example it is suggested that 15 tonnes of CO_2 are emitted for every tonne of lithium mined (Crawford 2022). The key points here are that there may be fundamental difficulties in meeting some of the SDGs and that the range of concerns for the 'doughnut economy' implicitly involves some conflicts. Thus, a complete argument needs to encompass the effects of the supply chain, including mining metals and rare earths, not just the fundamentals of human life such as air and water. Put another way, as well as considering overall goals we need to be mindful of the constraints in meeting those goals and their interdependencies which we will need to understand and respect when making trade-offs.

3 Framing the argument

The discussion above is mainly on a broad scale, e.g. concerns affecting nations and, in some cases, relates to impacts on a planetary scale. So, what do engineers do for a particular system of concern? Systems engineers normally consider trade-offs across a range of factors – although not as broad as we are proposing here. This raises questions of skills and competence. There are also questions about the extent to which decisions will be taken at a political level. These are complex issues which we can only partially address, but we return to them in the conclusions.

Here, we focus onto those concerns that: a) the system design can impact, and b) are in the scope of control or influence when designing the system. This enables us to establish an appropriate framing of the concerns. We approach this by first considering how to govern the safety of complex systems and second how to articulate arguments about the (ethical) acceptability of an individual system.

3.1 Safety of complex systems

The SDGs and many of the concerns in the 'doughnut economy' are at very broad scale and not something that can be controlled by any one system or development. However, if these concerns are not considered in developing or deploying individual systems then it is very unlikely that these goals will be met. A study of complex systems which exhibit emergent properties, e.g. effects of CO_2 emissions on the environment, proposed identifying safety controls at three layers – task & technical, management and governance, see Figure 5 (Burton et al 2021). We can broaden this model to cover safe, ethical and sustainable behaviour – and identify what concerns can and should be addressed at each layer.

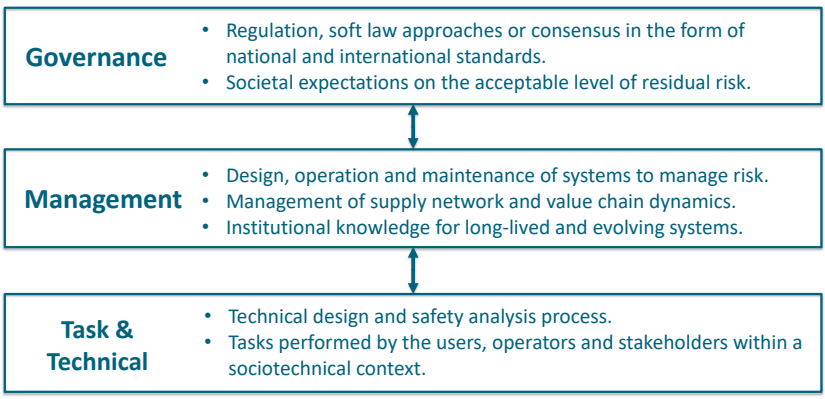

Fig. 5. Safety Management Layers (Burton et al 2021)

The critical decision-making that shapes systems and their wider impacts take place at the *management* layer – choice of system concept (and then the architecture, materials, capability, etc.) can all have a very broad effect. But an individual company making, say, an EV has no control over how electricity is generated (for their production or for individual vehicle owners), they do not dictate the policies for where rare earths are mined, etc. Thus, *governance* (let's say at the level of a nation) must put in place laws, policies, incentives, etc. that shape managerial decision-making. For example, this might be in terms of a maximum life-time carbon footprint for an EV together with policies for recycling critical materials. In this context, work at the *task & technical* layer mainly provides information on which management level decisions can be made, see the discussion on ethical assurance arguments below and the illustrative example in section 4.

3.2 Ethical assurance arguments

The use of safety cases is well-established, if not always well-practised. We have previously proposed extending the notion of safety cases to ethics – thus producing ethical assurance arguments (Porter et al 2022).

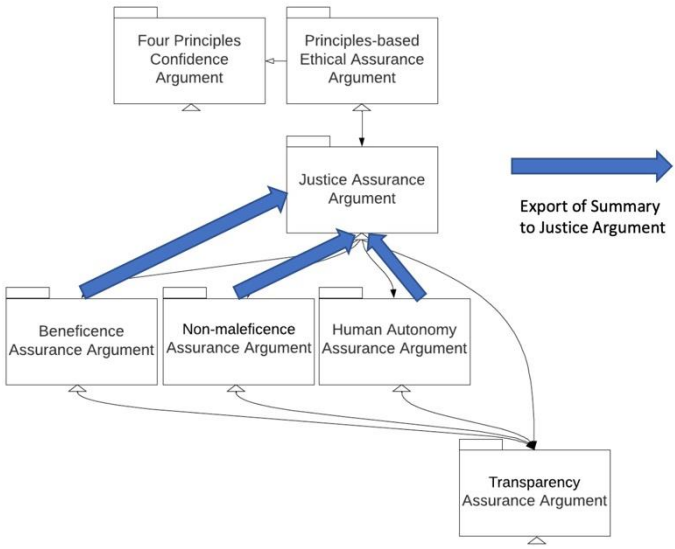

Fig. 6. Modular Structure of the Ethical Assurance Argument (Porter et al 2022)

The ethical assurance argument (see (Porter 2022) for more details) builds on concepts from biomedical ethics, as introduced above, in particular addressing:

- Beneficence – providing benefit from the system
- Non-maleficence – freedom from harm from the system (this includes safety considerations, i.e. avoidance of harm to individuals)
- Human autonomy – the ability to have appropriate human control over the system, including the ability to make meaningful choices about being affected by it, e.g. to opt out of its use
- Justice – the balance of benefit and risk from the system, including any constraints on human autonomy, across stakeholders, e.g. that benefits do not accrue to one set of stakeholders with the (risk of) harms falling on a different (disjoint) set of stakeholders.

As the concern in developing this argument structure was with AI/ML-based systems, and as ML models are often highly complex 'black boxes', demonstrating that the goals above have been met will often require the use of techniques that provide algorithmic visibility. Transparency is a supporting principle in the ethical assurance argument covering the ML models, and transparency of the assurance argument itself (i.e. visibility of the reasons and evidence for specific claims).

In using the argument framework, summaries of the argument in each of the four lower-level argument modules are exported to the justice argument. The approach can be adapted to deal with the wider safety, ethical and environmental concerns – safety is already addressed under non-maleficence and environmental concerns can be considered as both benefits and harms, e.g. removing plastic from the oceans is a benefit, and pollution of the atmosphere is clearly a harm. The scope of human autonomy might also be expanded – for example, in many cases nations make choices on behalf of citizens, e.g. on the mix of fuels for electricity generation, with individuals limited to narrower choices, e.g. choice of energy supplier.

4 An illustration – mobility as a service

To make the above ideas more concrete we consider the introduction of mobility as a service (MaaS) into a city, focusing on the decision-making at the level of the city (management layer). We assume that the city's current public transport system uses diesel-powered buses, and the Council wants to move to providing services autonomously, using EVs to improve air quality. The illustration shows how different options (vehicle mixes) can be compared using tables to summarise the beneficence, non-maleficence and human autonomy argument modules, so that the best approach can be chosen in the justice module. It is necessarily hypothetical but uses data, e.g. on vehicle carbon footprint, which is as realistic as possible. The illustration focuses on the trade-offs but see sections 4.8 and 4.9 for a discussion of the argument.

4.1 The status quo and the ambition

The city currently has a fleet of 50 buses (single and double-decker) providing about 250,000 passenger kilometres (pa-km) of journeys a day. The buses operate for 16 hours a day and travel is free for citizens over 60 years of age, after 9am. The city centre has several pedestrianised areas, and the Council has already limited access for private vehicles except to blue-badge holders (see below for an explanation) and residents. Unusually, the city doesn't allow taxis or private-hire vehicles to operate in the city centre[4].

The Council wishes to improve air quality and to provide transport through an MaaS scheme using EVs and asks their transport planners to come up with a range of options so they can choose the favoured approach. In doing the analysis to present ideas to the Council the transport planners decide that they need to consider different stakeholders, focusing on the users of the transport system:

[4] Not very likely but this can be viewed as an assumption to make the illustration more compact!

- Lone traveller – an individual who would prefer to travel alone (i.e. in a sole occupancy vehicle) for their safety and peace of mind
- Large family – parents and children travelling together who would prefer not to be separated[5]
- Young family – parents (or single parent) with one or two children using a pram or pushchair
- Blue badge – individuals with mobility needs and who are allowed to travel and park in areas otherwise not accessible to traffic
- Free travel – individuals living in the city and over the age of 60 entitled to a pass giving them free travel on the city's buses (after 9am)

There are also other stakeholders, including bus drivers and vehicle maintenance staff, whose employment can be affected by the MaaS operation.

4.2 Fleet options

The Council officials decide to explore three different vehicle mixes which they can then compare – including comparing against the current services.

Table 1. Fleet Options: Numbers of each type of vehicle

Vehicle Type	Fleet A	Fleet B	Fleet C
2 seat LPV	250	125	0
4 seat LPV	250	125	0
Bus	0	25	0
Shuttle	0	0	175

Council staff have seen several electric light passenger vehicles (LPVs), for example see Figure 7, and decide that one possibility (Fleet A) would be to use a mix of 2 and 4 seat LPVs. This would enable provision of services to lone individuals without having poor utilisation in larger vehicles, and the LPVs would be available on demand and able to go to a destination pre-selected by the passenger(s). The 4-seat LPVs are flexible in use and can be configured easily, for a particular journey, e.g. to hold a 2-seat buggy in the back (by folding down the rear seats) whilst still accommodating two adults in the front and thus are suitable for young families.

[5] In the UK around 28% of families have 3 or more children.

Fig. 7. Sven LPV[6]

A hybrid (Fleet B) would include buses on fixed routes, as now, along with a mix of 2 and 4 seat LPVs. The buses would only operate peak hours (8am to 6pm) and have a capacity of 60. For cost reasons, only half the number of LPVs is possible by comparison with Fleet A, but analysis of the current service usage indicates that this would cater for the city's needs as the LPVs would be available 24/7.

A third option is to use shuttles – like a mini-bus – with a maximum occupancy of 8 passengers (Fleet C) which can meet the overall transportation need but only by having journeys shared between different passenger groups.

4.3 Safety (non-maleficence)

The information that needs to be made available for scrutiny to support the justice argument was outlined in section 3.2 and needs to be at whole-fleet level. The data arising from traditional safety assessments can be summarised in two tables, where the percentages in Table 2 reflect the scores against specific vehicle tests as part of the Euro New Car Assessment Programme (NCAP)[7]:

Table 2. Euro NCAP Summaries by Vehicle Type

Vehicle Type	Adult Occupant	Child Occupant	Vulnerable Road Users	Safety Assist
2 Seat LPV	35%	42%	56%	24%
4 Seat LPV	65%	72%	60%	35%
Shuttle	91%	81%	73%	81%
Bus	23%	38%	18%	15%

[6] See: https://www.fev.com/en/media-center/press/press-releases/news-article/article/sven-shared-vehicle-electric-native-optimized-for-urban-mobility.html (accessed 16th October 2022)

[7] See: https://www.euroncap.com/en (accessed 2nd November 2022)

Table 3. Safety Ranking by Fleet

Safety	Fleet A	Fleet B	Fleet C
Rank	2	3	1

Buses perform poorly because of their mass (and shape) in impact with vulnerable road users (VRUs). The shuttles are best, as they are full-size vehicles and can be engineered with crumple zones, safety assistance systems such as automated emergency braking, etc. The LPVs are relatively poor as they don't have the size or power to include safety assistance features, with 2-seat LPVs worse due to limited space for crumple zones. The VRU score is quite high due to the low vehicle mass.

4.4 Environmental impact (non-maleficence)

The environmental impact is estimated[8] based on the full life of the vehicles including manufacturing as well as the operation of the vehicles; this implies some assumptions about how the electricity is generated as well as the useful vehicle life.

Table 4. Life-cycle carbon footprint of each vehicle type

Impact	2 Seat LPV	4 Seat LPV	Shuttle	Bus
Vehicle	60 g/CO_2/km	90 g/CO_2/km	230 g/CO_2/km	3,000 g/CO_2/km
Occupancy	1.5	2	5	20
Per Passenger	45 g/CO_2/pa-km	45 g/CO_2/pa-km	46 g/CO_2/pa-km	150 g/CO_2/pa-km

Table 5. Life-cycle carbon footprint for each fleet, for each day of use

Impact	Fleet A	Fleet B	Fleet C
LPV pa-km	250,000	150,000	0
Bus pa-km	0	100,000	0
Shuttle pa-km	0	0	250,000
Total (kg)	11,250	21,750	11,500

Given the data in table 4 including estimates of average occupancy, the carbon footprint of the different fleet options can be assessed, see table 5, setting out carbon cost per day. Fleet C is little different from Fleet A – although the shuttles are worse than the LPVs per vehicle kilometre this is more-or-less exactly balanced out by the higher average occupancy. The buses have a much worse impact due to their size and the expected occupancy.

[8] The figures are loosely based on an analysis of LPVs by Zemo (Zemo 2021).

4.5 Availability of transport (beneficence)

The primary intended benefit of the MaaS is to provide (better) transport services to citizens and visitors to the city. The benefits are assessed qualitatively with N meaning there is no substantial difference from the current bus service with +/++ and -/-- denoting smaller/larger improvements and detriments, respectively.

Table 6. Availability of transport per stakeholder for each fleet

Stakeholder	Fleet A	Fleet B	Fleet C
Lone traveller	++	+	N
Large family	--	-	+
Young family	-	-	+
Blue badge	--	--	N
Free travel	++	++	+

The table reflects value judgements – for example that Fleet A has problems for young families as they won't be able to get buggies into half of the LPVs. Fleet B is judged to be similar for young families as buses (which can easily take buggies) are on fixed routes and there are fewer 4 seat LPVs available. And so on.

4.6 Availability of employment (beneficence)

The employment benefits relate both to the bus drivers and to maintenance staff. Arguably the loss of employment could be seen as a harm, but we treat employment as a benefit for the purpose of this illustration.

Table 7. Availability of employment per stakeholder for each fleet

Stakeholder	Fleet A	Fleet B	Fleet C
Bus drivers	--	-	--
Maintenance	++	+	+

There may be compensation for the bus drivers, but that is left to the discussion of the argument.

4.7 Human autonomy

Human autonomy relates to the meaningful control related to the system, including appropriate freedom of choice for each stakeholder, for example the ability of a lone

traveller to ride in a single occupancy vehicle. In making the value judgements here we are assuming a level of control, for example being able to ask for a 2-seat LPV when calling for a vehicle. In the case of large and young families, the score reflects the need to wait for larger vehicles and the fact that very large families (5 or more individuals) would have to split up on a journey (with Fleet B, this would only be necessary outside core hours or if travelling far from the bus routes). Although blue badge holders, e.g. wheelchair users, can use buses and the shuttles, they would have less freedom – not being able to travel when they want and where they want, hence all the options represent a detriment.

Table 8. Autonomy per stakeholder for each fleet

Stakeholder	Fleet A	Fleet B	Fleet C
Lone traveller	++	+	-
Large family	--	-	+
Young family	--	-	N
Blue badge	--	-	-
Free travel	N	N	N

4.8 Making the argument

Ethical assurance arguments would have to be made for each concern – and sections 4.3 to 4.7 have illustrated the core data related to beneficence, non-maleficence, and human autonomy via the tables "exported" to the justice argument. In practice the arguments would be more complex, and the above should be viewed as fragmentary illustrations. Further, the arguments and evidence are subject to uncertainty and there will be a need to monitor systems in operation to collect leading indicators of risk, etc. providing continual assurance. We return to this point in the conclusions.

It is helpful to briefly consider the main argument modules before considering the justice argument. No consideration is given to transparency; it may be that these autonomous EVs use ML so transparency might be needed, for example, to estimate the VRU rating for the different vehicles. However, we view discussion of transparency, e.g. via explainability (McDermid et al 2021), as outside the scope of consideration here. There is also a wider issue of transparency of the assurance argument and evidence which we return to in the conclusions.

At the level of the three main argument modules – beneficence, non-maleficence, and human autonomy – there are two key criteria to meet. First, there are criteria which relate to the broad set of concerns identified above, e.g. the SDGs or elements from the (models of) the doughnut economy which "flow down" from the governance layer. This might, for example, be a requirement that some system is carbon neutral over its life – or, more likely in the case of the EVs in our illustration, there is some maximum carbon footprint per passenger kilometre based on knowledge

that there will be a carbon offset elsewhere. If, for example, the limit was 50g/CO_2/pa-km then Fleet B would be rejected. However, at a figure of 100g/CO_2/pa-km for the fleet then Fleet B would still be open for consideration even though the buses don't meet this target.

Second, there are internal criteria within each of the modules, for example, in the human autonomy argument, one such criterion would be that individual stakeholders can give informed consent to the use of the system in a way that affects them – for example, can a lone traveller choose who gets into a shared shuttle with them and/or the route the shuttle takes in completing its journey? A judgement might be made that, without such controls, Fleet C is unacceptable due to the constraints on the person's capacity to manage their own personal safety – especially late at night.

In both these cases, this can be viewed as being like an ALARP criterion – if a threshold is met then the option can be considered, in a similar way to being in the broadly tolerable region of ALARP. Those options that "survive" the module- level can be considered in the justice argument. In the illustration here, we assume that all three options "survive", so tables 3, 5, 6, 7 and 8 are the summaries that are "exported" to the justice argument as indicated in Figure 6.

Arguably, the ideal situation is that one of the options Pareto-dominates all the others, i.e. is better with regards to all the concerns. This isn't true in this case. Fleet C is, on most measures, as good as or better than the others (the difference in carbon footprint in table 5 is small enough to ignore) but there are significant disadvantages for bus drivers (table 7) and issues for both blue badge holders and lone travellers (table 8). There are potential resolutions (mitigations) for these issues across the safety management layers introduced above, e.g.

- Bus drivers – management layer change, offering retraining for other roles
- Blue badge holders – governance layer change, altering the policy to allow blue badge holders to enter the city centre as before (perhaps with incentives to adopt EVs)
- Lone travellers – management/task & technical layer change through adopting a different fleet mix (shuttles plus some LPVs) giving options for single occupancy travel (there are also single occupancy vehicles[9])

The approach to ethical arguments (Porter et al 2022) incorporates the notion of using (wider) 'reflective equilibrium' to reach a balance between conflicting demands – put another way, justifying the trade-offs. Reflective equilibrium is most closely associated with the work of the political philosopher John Rawls (Rawls 1951; Rawls 1971). In this context, we take wide reflective equilibrium to mean the end-point of a decision process which involves stakeholders (or their trusted representatives) and other decision-makers working back and forth between their considered ethical judgements about specific competing demands, general ethical principles that apply, and relevant non-ethical judgements (e.g. technical or financial) until they reach a coherent opinion. Reflective equilibrium is achieved when none of the parties involved are inclined to revise any of their component judgements or

[9] e.g. the Electra Meccanica Solo https://www.emvauto.com/solo (accessed 16th October 2022).

beliefs about the decision or trade-off further because together these have the highest degree of acceptability (Daniels 2020).[10] In this illustration, that approach might be used at two stages – initially identifying issues that can't be resolved, and later assessing (and accepting) the revised fleet choice and governance changes. One way of representing this would be to record the positions of each stakeholder.

4.9 Observations

One of the challenges for the 'safe, ethical and sustainable' mantra is the wide range of concerns (e.g. from the SDGs and the doughnut economy) that might need to be considered. Based on the illustration outlined above it seems reasonable to view the principles underlying the ethical assurance arguments as a good way of structuring concerns, whereas the SDGs and doughnut economy give good prompts, but neither can be viewed as exhaustive. Consequently, there is an important role for 'reflective equilibrium' in framing the argument for each system being considered although it remains to be seen how best to represent that in the justice argument.

However, in the discussion above there has been little explicit reference to framing the SDGs or the 'doughnut economy'. In effect, this framing "falls out" from the nature of the system (or at least it does, up to a point). We can consider each of the main elements of the argument in turn and how they relate to the broader goals:

- Non-maleficence (safety) – a specific aspect of SDG 3 (good health and well-being) but not something that is obviously reflected in the doughnut economy
- Non-maleficence (environment) – SDGs 11-13 (sustainable cities and communities, responsible consumption and innovation, and climate action) plus climate change and air pollution from the doughnut economy
- Benefits (employment) – SDG 8 (decent work and economic growth) and income and work from the doughnut economy
- Benefits (availability of transport) – not directly reflected although it arguably underpins employment benefits
- Human autonomy – perhaps this is implicit in gender equality (SDG 5) and reduced inequalities (SDG 10) and similarly maybe an aspect of social equality from the doughnut economy, but it is not really explicit

In addition, the layering of controls into governance, management and task & technical, enables assignment of responsibility for managing the concerns amongst different stakeholders who should have the appropriate knowledge and authority to discharge those responsibilities. This has been at least partially illustrated above.

[10] For good discussions of the method of reflective equilibrium in engineering ethics, see van de Poel & Swart (2010) and van den Hoven (1997)

5 Conclusions

The idea of considering whether systems are 'safe, ethical and sustainable' might be compelling in principle, but it is not so obvious how to meet this goal. The aim in this paper has been to shed some light on how this goal might be met by drawing together ideas from sustainable development, 21st century economics, management of safety in complex systems and ethical assurance arguments.

Our view remains that the key ethical concepts of beneficence, non-maleficence, human autonomy, and justice are a good way of structuring arguments about the acceptability of systems. Whilst we have not illustrated the dynamics of a reflective equilibrium process, it seems one of the few practical "tools" to address the trade-offs necessary between incommensurable concerns. Further, the layering of controls into task & technical, management and governance seems to help identify the best locus for addressing some of the overarching concerns.

However, the scope of potential concerns is vast and even the simple illustration here shows that the widely accepted SDGs and other models such as the doughnut economy do not embrace all the concerns – in particular, neither seem to reflect the notion of human autonomy very clearly or directly. Thus, we would advocate using the argumentation approach outlined here, treating the SDGs, doughnut economics – and potentially other frameworks – as checklists to make sure concerns which are important for a given system have not been overlooked. But our expectation is that, in many cases, the framing will "fall out" from appropriate consideration of the conceptual design for a system as it did in the illustration here. Time will tell if this is a realistic expectation in more complex settings.

Nonetheless, the analysis laid out above will, in reality, be much more complex, in particular as it will be based on evidence and subjective judgement that may include a high level of uncertainty and whose validity may erode over time as the environment in which the system operates, including behavioural patterns of the users, evolves. Thus, there will be a need for "continuous assurance" including an identification of the observation points required to collect leading risk indicators such that the system can be adapted to meet evolving needs and safety, ethical and sustainability trade-offs – which might also vary across countries and cultures.

But who has the skills and authority to conduct this work? Systems engineers are already used to designing systems to meet multiple, often conflicting criteria (e.g. performance, vs. cost vs. safety vs. usability). However, it is too much to expect that a technically trained engineer would have both the competence as well as the responsibility to handle this multitude of additional concerns. This is a place where reflective equilibrium has a role to play. Such a process can involve specialists who can represent these different specialisms. There is a need for systems (or safety) engineers who are capable of co-ordinating across the relevant disciplines. It is not uncommon to talk about a 'T-shaped' engineer who has broad skills and depth in one specific area, giving them authority (from the depth) and the skills to manage a multi-disciplinary team (from the breadth). The 'safe, ethical and sustainable' man-

tra suggests that the breadth for some engineers needs to be into ethical and environmental concerns, not just engineering issues. This implies a need for a refined education for at least a small cadre of engineers who have leadership roles in complex projects – and perhaps they need to be 'Π-shaped' with deep skills in two areas, perhaps environmental or ethical issues to complement a technical skill.

Further, this is where the layered model of (responsibility for) safety controls comes in. The issues should be considered from a system engineering and assurance perspective but at the management and governance layers where increasing responsibility is taken, and where this responsibility includes ensuring that the layer below produces systems that support the manifold safety, ethical and sustainability goals. This implies, for example, that the governance layer should produce an ethical assurance case, e.g. for the regulation of transport systems[11]. In other words, in order to deploy advanced technologies in a truly safe, ethical and sustainable manner, their deployment and regulation needs to be consciously "engineered" against a set of clear principles and methods for achieving transparency, including of the assurance arguments[12].

Finally, this leads to the question whether or not this approach would be broadly accepted by society used to elected politicians making "popular" decisions and in a time where rational, expert judgement is often not accepted and indeed rejected as "elitism". A sensitivity to these issues is therefore required, which is where both an ethical framing and understanding of societally perceived risk is essential. It is also one of the main motivations for the idea of a "mantra" (McDermid 2022) which, if repeated often enough, might begin to shape public and political perception and behaviours.

Acknowledgments This work is supported in part by the Assuring Autonomy International Programme, funded by the Lloyds Register Foundation, and the UKRI Trusted Autonomous Systems (TAS) programme through the Assuring Responsibility for TAS (AR-TAS) project.

References

Adams D (1982) Life, the universe and everything, Pan Macmillan.

Beauchamp T, Childress J (1979) Principles of biomedical ethics, Oxford University Press

Burton S, McDermid JA, Garnett P, Weaver R. (2021) Safer Complex Systems: An Initial Framework, https://raeng.org.uk/media/4wxiazh3/engineering-x-safer-complex-systems-an-initial-framework-report-v22.pdf (accessed 16th October 2022)

[11] This is not entirely unprecedented. For example, a safety case was produced for the change in vertical separation minima in European air space. What we are suggesting is a stretch beyond this, but the proposal is not without some antecedents.

[12] As an example of how this might be done in an "engineered" regulatory framework, the Centre for Data Ethics and Innovation has suggested that the Safe and Ethical Operating Concept for an autonomous road vehicle should be made publicly available, see: https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles/responsible-innovation-in-self-driving-vehicles#annex-a-safe-and-ethical-operational-concept-and-safety-management-systems.

Crawford I, (2022) How much CO_2 is emitted by manufacturing batteries? https://climate.mit.edu/ask-mit/how-much-co2-emitted-manufacturing-batteries (accessed 16th October 2022)

Daniels, N (2020) "Reflective Equilibrium", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/ (accessed 23rd October 2022)

Fjeld J. Achten N, Hilligoss H. Nagy, A, Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1).

Floridi L. and Cowls J. 2021. A unified framework of five principles for AI in society. In: Floridi L. (ed.) Ethics, Governance, and Policies in Artificial Intelligence. Philosophical Studies Series, 144: 81-90. Springer, Cham

Jobin, A, Ienca, M, Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9): 389-399

McDermid JA, Jia Y, Porter Z, Habli I (2021) Artificial intelligence explainability: the technical and ethical dimensions. Philosophical Transactions of the Royal Society A. 379(2207): 20200363.

McDermid JA, Porter Z, Jia Y (2022) Consumerism, Contradictions, Counterfactuals: Shaping the Evolution of Safety Engineering, in Safer Systems: The Next 30 Years, Proceedings of the 30th Safety-Critical Systems Symposium, Safety Critical Systems Club

McDermid JA. Safe, Ethical & Sustainable: A Mantra for All Seasons? (2022) Safety Systems. 2022 Feb 10:5-10.

Michaux SP (2021) Assessment of the Extra Capacity Required of Alternative Energy Electrical Power Systems to Completely Replace Fossil Fuels, Geological Survey of Finland, Report 42/2021

Michaux SP (2022) Private communication

Porter Z, Habli I, McDermid JA (2022) A Principle-based Ethical Assurance Argument for AI and Autonomous Systems. arXiv preprint arXiv:2203.15370.

Rawls, J., (1951) Outline of a decision procedure for ethics. The Philosophical Review, 60(2):177-197

Rawls, J. (1971) A theory of justice. Harvard University Press

Raworth K (2017) Doughnut economics: seven ways to think like a 21st-century economist. Chelsea Green Publishing

UN (2022) Sustainable development goals, https://www.un.org/sustainabledevelopment/ (accessed 16th October 2022)

Van den Hoven, J. (1997) Computer ethics and moral methodology. Metaphilosophy, 28(3): 234-248.

Van de Poel, I, Zwart, S.D. (2010) Reflective equilibrium in R & D networks. Science, Technology, & Human Values, 35(2): 174-199

Wiedmann T, Lenzen M, Keyßer LT, Steinberger JK (2020) Scientists' warning on affluence. Nature communications;11(1):1-0.

Zemo (2021) Powered Light Vehicles Life Cycle Analysis Study, https://www.zemo.org.uk/news-events/news,powered-light-vehicles-can-enable-transport-decarbonisationlifecycle-analys_4329.htm (accessed 16th October 2022)