# Autonomous military systems beyond human control

## Putting an empirical perspective on value trade-offs for autonomous systems design in the military

Boshuijzen-van Burken, C.G.; de Vries, M.O.; Allen, Janna; Spruit, S.; Mouter, N.; Munyasya, Aylin

**DOI**
10.1007/s00146-024-02000-3

**Publication date**
2024

**Document Version**
Final published version

**Published in**
AI and Society

AI & SOCIETY (2025) 40:2507–2523
https://doi.org/10.1007/s00146-024-02000-3

OPEN FORUM

# Autonomous military systems beyond human control: putting an empirical perspective on value trade-offs for autonomous systems design in the military

**Christine Boshuijzen-van Burken[1] · Martijn de Vries[2] · Jenna Allen[1] · Shannon Spruit[3] · Niek Mouter[2] · Aylin Munyasya[3]**

Received: 5 March 2024 / Accepted: 14 June 2024 / Published online: 7 July 2024


**Abstract**
The question of human control is a key concern in autonomous military systems debates. Our research qualitatively and quantitatively investigates values and concerns of the general public, as they relate to autonomous military systems, with particular attention to the value of human control. Using participatory value evaluation (PVE), we consulted 1980 Australians about which values matter in relation to two specific technologies: an autonomous minesweeping submarine and an autonomous drone that can drop bombs. Based on value sensitive design, participants were tasked to enhance the systems with design features that can realize values. A restriction (limited budget) in each design task forced participants to make trade-offs between design options and the values that these options realize. Our results suggest that the 'general public' has diverse and nuanced stances on the question of human control over autonomous military systems. A third of participants that is opposed to autonomous military systems when asked directly, selected different combinations of design features realizing varying degrees of human control. Several contextual factors, technology-specific concerns, and certain values seemed to explain these different choices. Our research shows that a focus on human control might overlook other important values that the general public is concerned about, such as system reliability, verifiability, and retrievability.

**Keywords** Autonomous military systems · Value sensitive design · Participatory value evaluation · Human control · Latent class cluster analysis

## 1 Introduction

Ukrainian unmanned submarines strapped with explosives approached the Russian naval fleet in an attempt to stop them from launching cruise missiles towards Ukrainian cities, but then they "lost connectivity and washed ashore harmlessly", writes Isaacson in his recent biography of Elon Musk (Isaacson 2023). To disrupt a Ukrainian sneak attack on the Russian fleet, Elon Musk told his engineers not to turn on his company's Starlink satellite communications network near the Crimean coast in 2022. The effect was that human control over the submarines was lost.

The question of human control is one of the key issues in many legal and ethical debates on autonomous systems that can be used for military purposes. Human control has been discussed at the United Nations (UN) Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) in discussions of emerging technologies in the area of lethal autonomous weapons systems (LAWS) since their inception in 2013. Activists, legal experts, academics, and international spokespeople continue to discuss the importance of human control. Some call for a total ban on the development of autonomous military systems, arguing that they will inevitably take action with severe consequences when beyond human control (Asaro 2012; Docherty 2020; Future of Life Institute 2015; Russell 2023), while others demand meaningful human control over the decisions that autonomous systems make (Amoroso and Tamburrini 2020; Watch et al. 2016; Santoni de Sio and Hoven 2018; Steen et al. 2023). What exactly makes up human control is not agreed upon and perceptions vary greatly, from narrow

✉ Christine Boshuijzen-van Burken
c.vanburken@unsw.edu.au

[1] The University of New South Wales, Canberra, Australia

[2] TU Delft, Delft, The Netherlands

[3] Populytics, Leiden, The Netherlands

understandings that focus on the operator or commander, to broad understandings, such as Australia's contribution to the UN CCW debate in 2019. They pointed to "Australia's System of Control and applications for Autonomous Weapon Systems" (GGE LAW 2019), to show that humans make decisions and thus exert control in various ways, for example by means of legal reviews (under Article 36 of International Humanitarian Law), testing and training of people and military systems.

In 2023, the UN Secretary-General and the President of the International Committee of the Red Cross (ICRC) jointly called on world leaders to "act now to preserve human control over the use of force" (United Nations 2023). Finally, implicit references to human control can be found in the principles and guidelines that have emerged over the last few years for the responsible use of artificial intelligence (AI) in the military, such as accountability, human centricity, controllability, and governability (Devitt et al. 2021; NATO 2021; DoD 2022, 2023). Many of these principles and others such as human dignity, autonomy, privacy, transparency, have been discussed in academic literature, including in this journal (Blanchard et al. 2024; Cebulla et al. 2023; Paraman and Anamalah 2023; Sadek et al. 2024). The question of human control over autonomous military systems is clearly important to many experts, but what is less clear is how the general public appreciates this highly debated topic. In this paper, we provide an in-depth empirical analysis of values related to autonomous military systems with special attention to the value of human control. We show that the general public values human control over autonomous military systems but that other values are equally or more important to them.

The field of responsible research and innovation (RRI) argues that it is important to embed values that citizens deem important in the design of technology to increase its acceptance (Owen, et al. 2013; Taebi et al. 2014). This holds even more true when designing autonomous systems for military purposes, as these technologies are used by the military on behalf of society (at least in liberal democracies where governments—who own the monopoly on the use of force—act on behalf of their citizens) in life-or-death situations.

Over the past decade, various surveys have been conducted, such as those done by the Campaign to Stop Killer Robots (Conboy 2021; Deeney 2019), in which the public was asked about their stance towards lethal autonomous weapons systems, defined as "weapons systems [that] would be capable of independently selecting targets and attacking those targets without human intervention" (ibid). This survey found opposition to the use of autonomous weapons and artificial intelligence in offensive military operations, which is in line with the findings of similar surveys (Horowitz 2016; Rosendorf et al. 2022; Zhang, et al. 2021). However, the choice of terminology in these surveys might importantly influence the answers respondents give. For instance, the prospect of 'fully autonomous' lethal systems as described by the Campaign to Stop Killer Robots has been regarded as science fiction that is void of a sense of reality in actual military operations (Conboy 2021; Bo 2022; Rosert and Sauer 2021; Young and Carpenter 2018). Other scholars added that this individualistic focus on human intervention during targeting, typically presented as an autonomous system replacing a combatant on the battlefield or the commander, is a simplification of actual military practice that may hinder meaningful discussions about human control (Ekelhof 2019, 2018; Verdiesen 2017). Hadlington and colleagues found evidence that the public has indeed misunderstandings about the use and implementation of AI in the military (Hadlington et al. 2024).

Other surveys measured sentiments or attitudes towards autonomous military systems or the ascriptions of responsibility for the use of these systems, and their methods typically employed a binary answer or Likert scale (Conboy 2021; Arai and Matsumoto 2023; Verdiesen et al. 2019; Verdiesen and Dignum 2022; Lillemäe et al. 2023). A downside of these set-ups is that value preferences are measured in a utopian and abstract manner since respondents do not see the consequences of their choices and do not have to make value trade-offs. It is well known that this could lead to hypothetical bias: respondents might overestimate the importance of their stated values, because in the abstract, they are always important (Carson and Groves 2007; Johnston et al. 2017). This may explain some conflicting results between studies (at least between Verdiesen and colleagues (Verdiesen et al. 2019) and Arai and Matsumoto who had opposite findings when participants were presented with an abstract scenario and asked about the use of LAWS (Arai and Matsumoto 2023)). Moreover, the absence of a constraint potentially limits the real-life usability of these studies for decision-makers, who must make choices subject to constraints.

These limitations can potentially be rectified by conducting a participatory value evaluation (PVE). PVE is a preference elicitation method originally developed to measure the societal value of government policies (Mouter et al. 2021a). The Dutch government deployed PVE to investigate citizens' preferences and shared values for COVID-19 policies (Mouter et al. 2022; Mouter et al. 2021b), transportation policies (Mouter et al. 2021c), environmental policies (Itten and Mouter 2022) and flood protection policies (Mouter et al. 2021b). Recently, the method has also been applied in Austria (Hössinger et al. 2023), Israel (Golan 2023) and Peru (Gonzales Pecho 2023).

The essence of a PVE is that participants are effectively placed in the seat of decision-makers. In an online environment, they (a) see which options the decision maker is considering, (b) consider the concrete impacts of the options, and (c) have to make choices within given constraint(s).

Subsequently, citizens are asked to provide a recommendation on the policy options the government should choose, subject to the constraint(s). Next, they are asked to explain in writing their reasoning/motivation for each of their choices. The main virtue of PVE is that respondents are forced to make value trade-offs, whereas in the Likert-scale experiments, respondents express their preferences without any constraint (Conboy 2021; Verdiesen et al. 2019; Verdiesen and Dignum 2022). Moreover, in a PVE, detailed information can be provided about the meaning and consequences of choices, adding to the realism and usability of the results for decision-makers.

In this paper, we use the PVE in a slightly different manner from the one described above, as we are not measuring government policy preferences. This PVE aims to inform participants about technical design choices to support a more inclusive technology design process. The main goal of this PVE is to investigate which values people would like to see embedded in two autonomous military systems: an autonomous submarine for mine countermeasures and an autonomous aerial drone that can drop bombs. We elicited value preferences from a representative sample (1980 participants) of the Australian population for the design features of these autonomous systems when participants are confronted with realistic design choices where value trade-offs have to be made.[1] To this end, we report the design choices participants make using a statistical clustering method, providing insights into how different subgroups of the populations make value trade-offs when designing certain technologies. We also analyse the written explanations that citizens provide for their choices to identify segments of the population who have opposing opinions, particularly about the value of human control. This is the second contribution of our study because existing empirical studies on the topic of autonomous systems for military purposes rarely include qualitative data analysis (the exception is Verdiesen and Dignum (2022)). This approach allows us to obtain a richer understanding of why people make certain value choices beyond any preliminary explanations that are based on demographics (e.g., age, educational level, attitudes towards defence). Third, design options are presented in an integral manner, including detailed information about their meaning and consequences. The final contribution is that this is the first time that the PVE method has been applied in Australia and the first time that the method has been applied to military technology design.

[1] The idea of implicitly or explicitly embedding values in technology is widely discussed in literature on value-sensitive design. (Boshuijzen-van Burken 2023) provides a case in point for autonomous military systems.

The organization of this paper is as follows. We start with relevant definitions for our topic. We then explain our PVE survey method where participants made value trade-offs while designing two autonomous military systems, namely, an autonomous submarine for mine countermeasures and an autonomous aerial drone that can drop bombs. The results section comprises a quantitative analysis that reveals a striking value pluralism and a qualitative analysis of our survey data that reveals the underlying concerns and additional values that were not captured in the design task. In the discussion section, we discuss interesting results, and we then close this paper with a conclusion and suggestions for further research.

## 2 Definitions

We adopt the following key definitions related to our topic, namely, 'autonomous systems' and 'weapon' and 'values'.

Definitions of 'autonomous systems' vary greatly, including definitions of autonomous systems for military use (Taddeo and Blanchard 2022). For the purposes of this paper, we consider autonomous systems to be systems that can perform actions or make decisions with little or no human intervention, often involving some degree of artificial intelligence (AI). Furthermore, we concur with Burton that "autonomous decision-making capability can be understood as a spectrum" (Burton et al. 2020), with human interference on one end. The idea of degrees of autonomy is reflected in the content of the PVE, as we provide design options that provide either no or great degrees of human control over the decisions and actions of the system.

To define a weapon, we adopt Australia's formal definition (Australian Government. Department of Defence 2021):

"1. Weapon: Any device, whether tangible or intangible, designed or intended to be used in warfare to cause:

a. Injury to or death of persons; or
b. Damage to, or destruction of, objects".

In short, devices that are intended to cause the injury or death of persons or damage or destroy objects in warfare should be thought of as weapons. Such weapons in the broad sense can be designed to act autonomously, and we included an example in our survey, namely, a submarine for minesweeping.

Last, for the purposes of this paper, we refer to Friedman, Kahn and Borning, who propose a definition of values that works well with value-sensitive design, which is the broader context in which our research stands. They state that a value is something that "[…] a person or group of people consider important in life" (Friedman et al. 2006, p. 349).

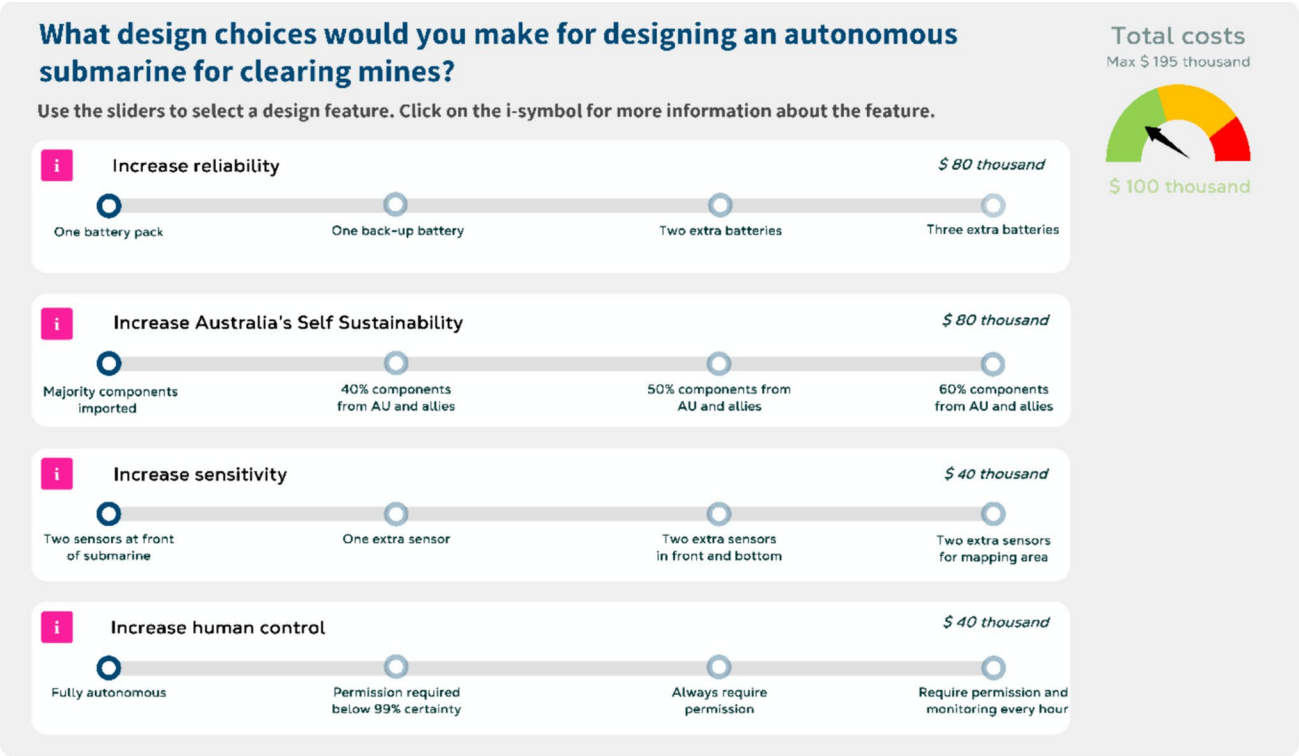


**Fig. 1** Screenshot of design task 1 for the autonomous submarine for minesweeping

## 3 Methodology

### 3.1 Participatory value evaluation

PVE is a value elicitation method for mapping values in a large and diverse group of citizens. It was originally designed as an economic appraisal method for assessing the societal value (aggregate utility) of government policy options (Mouter et al. 2019). This PVE aims to elicit information from participants in a value-based technical design process with specific attention to human control. The design options presented in this PVE all maximize a specific value. Participants make a value trade-off (they prioritize certain values over others) because they make choices subject to a constraint: the costs associated with various design options.

The PVE content for this survey was created through a codesign process with hardware and software engineers, legal and ethical scholars and representatives of various NGOs relevant to the topic, and this survey has been reported elsewhere (see Boshuijzen-van Burken et al. 2023). Via an online anonymous focus group, the diverse stakeholder group suggested an autonomous submarine for minesweeping and an autonomous aerial drone for dropping bombs to mitigate threats to Australia. The stakeholders also specified relevant values and a set of design features that can realize the suggested values. We adopted these for use in both PVE design tasks (Figs. 1 and 2). Participants could click on the "i" symbol next to each design feature for a brief explanation, including possible advantages and disadvantages of design features.

To force participants into value trade-offs, each design feature had a cost, and only a preset budget could be spent, allowing participants to choose at most approximately 25% of the total available options. Participants could not submit the survey if they had overspent their budget. The costs associated with the various design options were determined in consultation with subject-matter experts to provide participants with realistic monetary trade-offs.

At the start of the survey, we asked participants two questions that we used as a proxy for pro- or anti-defence sentiments: whether the government should spend more money on defence and whether people would support their sons or daughters joining the defence forces. We also asked people about their 'support for autonomous systems that can select and attack targets' to measure the general sentiment towards these systems and to capture any intrinsic antagonism that people may have towards these systems. These sentiments may partially provide reasons why people make certain decisions in the design task.

In the first PVE task, participants were asked to design an autonomous submarine for clearing underwater mines. Participants received information about the default design of

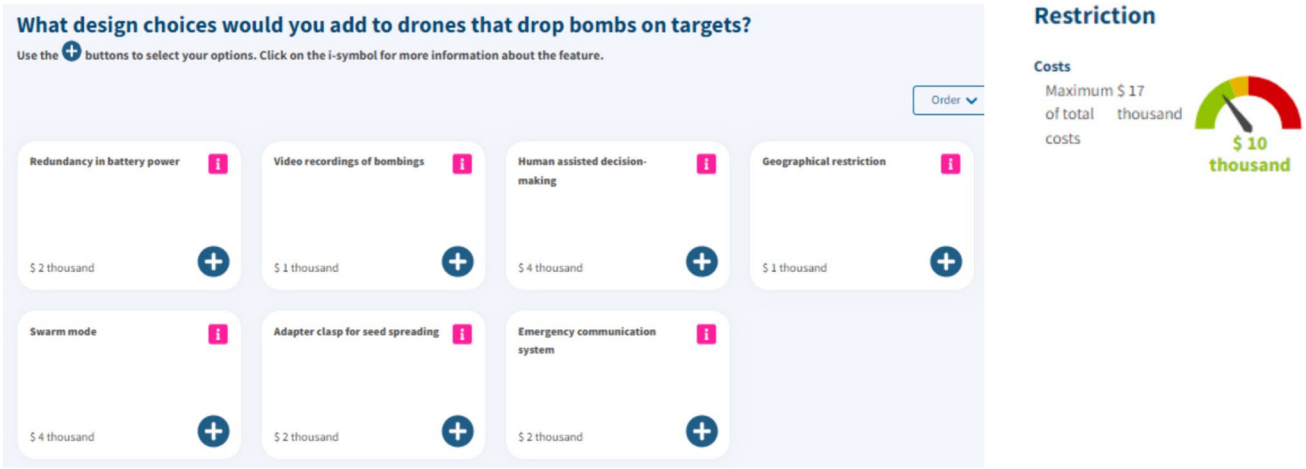


**Fig. 2** Screenshot of design task 2 for the armed autonomous drone

the submarine, for example, that the submarine can autonomously detonate mines if it is 99% sure that it has detected a mine. Participants were asked to make choices between four different design features that reflect values: (1) increase human-assisted decision making to increase human control, (2) add sonar sensors to promote sensitivity, (3) increase the use of Australia's supply chain to promote Australia's self-sustainability, and (4) increase the number of batteries in the autonomous system to promote system reliability. Participants enhanced the submarine with design features by moving a slider to the right. For each value (e.g., human control), participants could choose from four slider positions. Each slider position increases or decreases the realization of the value through the design feature. The more the slider moves to the right, the more the value is realized (see Fig. 1).

The second design task was an autonomous aerial drone that can drop bombs (see Fig. 2). Participants received information about the default design of the drone, for example, that the drone is trained to autonomously detect and strike targets with precision and that it had proven to accurately strike the intended target in 98% of cases. Participants simply chose a design feature (or not), so contrary to the slider mode of Design Task 1, there were no degrees or intensities by which a value would be realized. Participants were asked to make choices among seven different design features: (1) additional battery power, (2) video recordings of bombings, (3) human-assisted decision making, (4) geographical restriction, (5) a swarm mode, (6) an adapter clasp for seed spreading, and (7) an emergency communication system.

## 3.2 Data collection

Participants in the experiment were sampled from an online panel (Dynata), with a view to be representative of the Australian population in terms of age, gender and education. The Human Research Ethics Committee of UNSW approved our study protocol (HC220732). The experiment ran from 13 March 2023 to 22 May 2023, and a total of 1980 participants completed the questionnaire. The full list of questions is available upon request.

Table 1 gives an overview of the sociodemographic characteristics of the sample. Because some strata are slightly under- or overrepresented, the data have been weighted in all analyses for both surveys using poststratification weights. Based on the characteristics of gender (2 groups), age (7 groups) and highest education level attained (3 groups), the participants could be divided into 42 different strata. The relative size of each stratum was compared to that of the Australian population in 2022 (Australian Bureau of Statistics, 2022). The weight of each stratum was then calculated by dividing the proportion of the population by the proportion of the sample.

## 3.3 Statistical analysis: latent class cluster analysis

We first analysed the data using standard descriptive statistics. Subsequently, we analysed the choices participants made using a latent class cluster analysis (LCCA). This method is ideally suited to identify common patterns in the design choices made by different groups (clusters) of people. Based on maximum likelihood estimation, a model identifies clusters that are maximally homogeneous within clusters (consisting of people who make similar patterns of design choices) and maximally heterogeneous between the clusters.

A benefit of LCCA is that covariates can be included in the model to assess their associations with cluster membership. In so doing, the model can reveal which segments of the population (e.g., in terms of age or gender) relatively frequently belong to a certain cluster. This makes it possible to determine which (combinations of) design choices are

**Table 1** Sociodemographic characteristics of the sample

| Variable | Frequency | Proportion % |
|---|---|---|
| Total | 1.980 | 100 |
| Gender | | |
| Male | 928 | 47 |
| Female | 1.040 | 53 |
| Different or prefer not to say | 12 | 1 |
| Age | | |
| Younger than 20 | 96 | 5 |
| Between 21 and 30 | 373 | 19 |
| Between 31 and 40 | 410 | 21 |
| Between 41 and 50 | 342 | 17 |
| Between 51 and 60 | 314 | 16 |
| Between 61 and 70 | 258 | 13 |
| Older than 71 | 187 | 9 |
| Education | | |
| Incomplete secondary education | 170 | 9 |
| Secondary education completed | 365 | 18 |
| Some university or vocational certification | 382 | 19 |
| Vocational or professional certification completed | 264 | 13 |
| University education completed | 596 | 30 |
| Doctorate, postdoctorate or equivalent completed | 35 | 2 |
| Information not provided | 168 | 8 |

**Table 2** Model fit results of LCCA models for Design Task 1

| Design task | # Clusters | LL | BIC(LL) | Npar | $L^2$ | df | *p* value | Class.Err |
|---|---|---|---|---|---|---|---|---|
| Task 1 | 1-Cluster | –7740.21 | 15,571.15 | 12 | 3135.423 | 243 | 2.2e-496 | 0 |
| Task 1 | 2-Cluster | –7413.07 | 15,211.84 | 51 | 14,263.64 | 1874 | 5.2e-1868 | 0.0108 |
| Task 1 | 3-Cluster | –7168.38 | 15,017.38 | 90 | 13,774.25 | 1835 | 6.4e-1793 | 0.0462 |
| Task 1 | 4-Cluster | –6928.86 | 14,833.29 | 129 | 13,295.22 | 1796 | 7.5e-1720 | 0.0103 |
| Task 1 | 5-Cluster | –6814.29 | 14,899.09 | 168 | 13,066.08 | 1757 | 9.0e-1694 | 0.0554 |
| Task 1 | 6-Cluster | –6697.56 | 14,960.56 | 207 | 12,832.61 | 1718 | 7.0e-1667 | 0.062 |
| Task 2 | 2-Cluster | –7397.8002 | 15,166.16 | 49 | 14,201.82 | 1876 | 1.8e-1855 | 0.076 |
| Task 2 | 3-Cluster | –7227.305 | 15,142.8 | 91 | 13,860.83 | 1834 | 1.2e-1809 | 0.0588 |
| Task 2 | 4-Cluster | –7131.192 | 15,268.2 | 133 | 13,668.6 | 1792 | 6.6e-1792 | 0.1246 |
| Task 2 | 5-Cluster | –7051.836 | 15,427.12 | 175 | 13,509.89 | 1750 | 1.6e-1780 | 0.1331 |
| Task 2 | 6-Cluster | –6949.314 | 15,539.7 | 217 | 13,304.85 | 1708 | 2.0e-1760 | 0.1057 |

relatively frequently selected by certain groups of participants. This subgroup analysis can be used to identify conflict and consensus between subgroups of the population about values and design choices for autonomous military systems. The following covariates were considered in the analyses: gender, age, level of education, region, and opinion about autonomous systems in Australia, support son/daughter going into the military, support government expenditure on military, and perceived safety.

The aim of an LCCA is to find the most parsimonious model (with the least number of parameters) that adequately describes the associations between choices and covariates. To identify the optimal number of clusters for each design task, we estimated models with 1–6 classes using Latent Gold (Vermunt and Magidson 2013). Based on the Bayesian Information Criterion (BIC), the optimal model for Design Tasks 1 and 2 consisted of the 4-cluster and 3-cluster models, respectively (Table 2).

### 3.4 Qualitative data analysis

The qualitative analysis of the motivations participants provided for their design choices proceeded in two stages. First, two researchers independently open coded the first 300 participants' motivations, paying particular attention to the values and concerns that participants mention as motivation

**Table 3** Percentage of participants who (did not) support autonomous systems

| Autonomous systems are technologies that require little or no human intervention to operate. Which of these statements reflects your opinion on autonomous systems for Australia's defence? | Percentage % |
|---|---|
| I support the development of autonomous technologies that can detect and strike a target | 54 |
| I don't support the development of autonomous technologies that can detect and strike a target, unless they are better at detecting and striking targets than humans are | 25 |
| I don't support the development of autonomous technologies that can detect and strike a target, even if they are better at detecting and striking targets than humans are | 8 |
| I don't know | 13 |

for or against certain design features. To ensure intercoder reliability, both codebooks were compared. The majority of codes were exactly the same or similar, but some codes were unique to one of the researchers. After discussing the differences, a final codebook was made. This was used for closed coding the motivations of 700 participants, providing an overview of how frequently certain values and concerns were mentioned.

## 4 Results

### 4.1 Opinion about autonomous systems

Prior to the PVE design tasks, we asked participants several survey questions about their opinions on the military in general and autonomous technologies in particular. These survey questions were used as covariates in the LCCAs (see Sect. 3.3). Table 3 shows that 54% of the participants supported autonomous systems, whereas 33% did not.

### 4.2 Design task 1 for the autonomous minesweeping submarine

#### 4.2.1 Descriptive results

Figure 3 shows the share of participants that selected the various design options for the autonomous minesweeping submarine.

The value that was most often maximized by participants was Australia's self-sustainability. Twenty-five percent of participants designed an autonomous minesweeping submarine with at least 50% of the components coming from Australia and its allies. Of this group, 5% maximized this design requirement, so that (4) 60% of components are from Australia and its allies.

As the second most frequently maximized value, increasing sensitivity closely follows the choice to increase Australia's self-sustainability. Although the group that selected the highest (4) or second highest (3) design requirement for this value is smaller than the group that maximized the value of self-sustainability, a larger proportion of participants selected a slight increase in sensitivity (56%) than those who slightly increased Australia's self-sustainability (51%). Put differently, the share of participants who did not increase Australia's self-sustainability (24%) is larger than the share of participants who did not increase the device's sensitivity (22%).

The value that was the least often increased was human control. Thirty-one percent of all participants chose (1) to leave the device fully autonomous instead of increasing human control. That being said, 61% selected a slight increase in human control at (2): 'permission required below 99% certainty'. Only 6% selected (3): 'always require permission'. Only 1% maximized the value of human control at (4): 'require permission and monitoring every hour'.

#### 4.2.2 LCCA results

Table 4 shows the results of the LCCA for Design Task 1 for the autonomous minesweeping submarine. Only two of the included covariates (Sect. 3.3) were significant in the model: citizens' opinion on autonomous systems in Australia and whether they would support their son or daughter joining the military. The four design choices were included in the model as ordinal indicators: moving the slider position to the right (see Fig. 1) represents an increase in the value reflected by the design feature. Hence, the percentage scores of the indicators represent a probabilistic estimation of the average slider position selected by each cluster: 0% corresponds with position 1 (completely on the left), 100% with position 4 (completely on the right) (see Fig. 3).

The LCCA shows that most participants (Cluster 1, 69% of participants) consider all values to be somewhat important for the design of an autonomous minesweeping submarine (average score of 27–44% on a scale from 0% to 100%). Consistent with the descriptive results (Fig. 3), 'increasing Australia's self-sustainability' and 'increasing sensitivity' are the two highest ranked values in this cluster. Most participants in this cluster (55%) supported the development of autonomous technologies, and a majority supported their son or daughter going into the military (66%).

Interesting differences can be observed in the choices the other three clusters make. On the one hand, Cluster 2 (15%

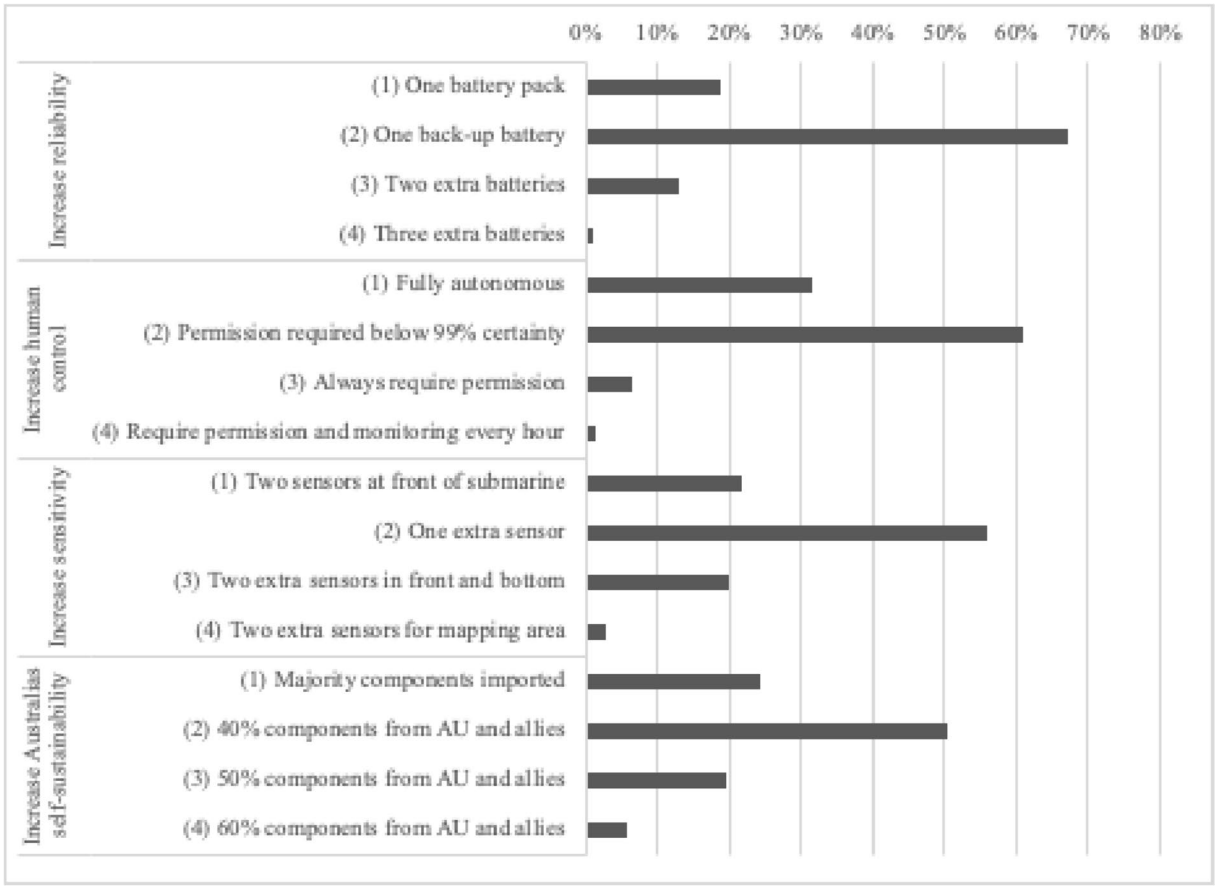


**Fig. 3** Distribution of design choices for the autonomous minesweeping submarine

**Table 4** Results of the LCCA for Design Task 1 (autonomous minesweeping submarine)

| | Cluster1 % | Cluster2 % | Cluster3 % | Cluster4 % |
|---|---|---|---|---|
| Cluster Size | 69 | 15 | 8 | 9 |
| Indicators | | | | |
| Increase reliability | 30 | 69 | 16 | 1 |
| Increase human control | 27 | 8 | 71 | 2 |
| Increase sensitivity | 40 | 38 | 17 | 3 |
| Increase Australia's self-sustainability | 43 | 28 | 18 | 1 |
| Covariates | | | | |
| Opinion on autonomous systems in Australia | | | | |
| I don't know | 11 | 11 | 13 | 23 |
| I don't support | 33 | 24 | 54 | 33 |
| I support | 55 | 65 | 33 | 44 |
| Would you support son/daughter going into military? | | | | |
| I don't know | 3 | 4 | 5 | 8 |
| I would suggest a different occupation | 17 | 12 | 18 | 11 |
| Neutral | 14 | 16 | 16 | 21 |
| I would support that step | 40 | 42 | 28 | 34 |
| I would strongly support that step | 26 | 27 | 33 | 26 |

**Table 5** Average design choices for the armed autonomous drone

| Design option | Proportion selected % |
|---|---|
| Redundancy in battery power | 32 |
| Video recordings of bombings | 40 |
| Human-assisted decision making | 28 |
| Geographical restriction | 26 |
| Swarm mode | 10 |
| Adapter clasp for seed spreading | 18 |
| Emergency communication system | 39 |

of participants) did not consider an increase in human control to be an important value for the design of an autonomous minesweeping submarine (average score of 8%). This is consistent with the result that most participants within this cluster supported the development of autonomous technologies (65%). Participants in Cluster 2 assigned the most value to 'increasing reliability' (average score of 69%). An interesting observation is that participants who assigned much value to this design requirement also frequently assigned substantial value to 'increasing sensitivity' (average score of 38%).

On the other hand, Cluster 3 (8% of participants) did consider the increase in human control to be the most important value for the design of the minesweeper (average score of 68%). This is consistent with the result that most participants within this cluster did not support the development of autonomous technologies (55%). Participants who assigned much value to increasing human control on average considered the other three values to be equally (un)important.

Finally, cluster 4 (9% of participants) did not increase any of the four design requirements. There can be several explanations for this; for example, this group may have had technical difficulties with moving the slider, or they were not interested in including any of the design options or were mostly nontraders (people who did not take the survey seriously).

## 4.3 Design task 2 for the armed autonomous drone

### 4.3.1 Descriptive results

Table 5 presents the average design choices for the armed autonomous drone. The three most frequently selected design options were 'video recordings of bombings' (selected by 40% of participants), the 'emergency communication system' (39%) and 'redundancy in battery power' (32%). The two design options that were least frequently selected were the 'adapter clasp for seed spreading' (18%) and 'swarm mode' (10%). Human-assisted decision making was selected by 28% of the participants.

Comparative analysis of Design Task 1 and Design Task 2 revealed that 23% of people who chose no human control in Design Task 1 for the submarine (slider to the left) chose human-assisted decision making for the drone in Design Task 2.

### 4.3.2 LCCA results

Table 6 show the outcome of the LCCA for Design Task 2. Three of the included covariates (Sect. 3.3.) were significant in the model: opinion on autonomous systems in Australia, support son/daughter going into the military and perceived safety. The seven design choices were included in the model as binominal indicators, as participants could choose to select the option or not (see Fig. 2). Hence, the percentage scores of the indicators correspond to the probability that respondents in that cluster selected the design choice: 0% means no one selected the option, and 100% means everyone selected the option. Some interesting patterns can be observed in Table 6.

In Cluster 1 (58% of participants), no one chose human-assisted decision making. Redundancy in battery power (45%), video recordings of bombings (42%) and geographic restriction (30%) are the most frequently selected design features by participants in this cluster. Most participants in this cluster (59%) supported autonomous systems when asked by a survey question (Table 3).

In Cluster 2 (23% respondents), everyone chose the emergency communication system (100%). In addition, human-assisted decision making (44%) and video recordings of bombings (34%) were most frequently selected by participants in this cluster. Half of the participants in this cluster (49%) supported autonomous military systems when asked directly.

In Cluster 3 (19% respondents), almost everyone chose human-assisted decision making (99%), and no one chose the emergency communication system (0%). Furthermore, participants in this cluster most frequently selected video recordings of bombings (38%), followed by redundancy in battery power (29%). This cluster had the least support for autonomous military technologies when asked directly (45%).

## 4.4 Qualitative analysis

In this section, we report the results of the qualitative analysis of the written motivations participants provided for their choices. In so doing, we pay particular attention to the values and concerns that motivate decisions for or against certain design features. We first discuss the results of each design task separately and subsequently discuss general insights. All quotations are unaltered from the written motivations

**Table 6** Results of the LCCA for Design Task 2 (armed autonomous drone)

| | Cluster 1 % | Cluster 2 % | Cluster 3 % |
|---|---|---|---|
| Cluster size | 58 | 23 | 19 |
| Indicators | | | |
| Redundancy in battery power | 45 | 0 | 29 |
| Video recordings of bombings | 42 | 34 | 38 |
| Human-assisted decision making | 0 | 42 | 99 |
| Geographical restriction | 30 | 16 | 23 |
| Swarm mode | 16 | 3 | 0 |
| Adapter clasp for seed spreading | 25 | 0 | 18 |
| Emergency communication system | 28 | 100 | 0 |
| Covariates | | | |
| Opinion on autonomous systems in Australia | | | |
| *I don't know* | 12 | 15 | 12 |
| *I don't support* | 30 | 36 | 42 |
| *I support* | 59 | 49 | 46 |
| Would you support son/daughter going into military? | | | |
| I don't know | 3 | 3 | 5 |
| I would suggest different occupation | 13 | 19 | 22 |
| Neutral | 17 | 10 | 14 |
| I would support | 40 | 41 | 34 |
| I would strongly support | 27 | 27 | 26 |
| Perceived safety | | | |
| I don't know | 1 | 2 | |
| Very unsafe | 5 | 7 | 5 |
| Unsafe | 20 | 25 | 18 |
| Neutral | 30 | 30 | 35 |
| Safe | 37 | 27 | 33 |
| Very safe | 7 | 9 | 6 |

submitted by anonymous PVE participants, unless stated otherwise.

#### 4.4.1 Design task 1: the autonomous minesweeping submarine

In the first design task, participants were presented with four values that they could increase by moving a slider to select the degree of application of a specific design features (Fig. 1). Below, we present the results of the qualitative analysis of the motivations people gave for their design choices.

Participants make selections with a concern for self-reliance.

Participants raised numerous arguments and concerns reflecting the attitude that Australia needs to take care of itself. These were especially prevalent in participants' decisions to increase Australia's self-sustainability. Australia's economy was often mentioned. People reasoned that Australia's self-sustainability "helps the work force and our industries". Moreover, some people expressed a lack of trust in overseas partners, preferring "not to rely on overseas sources/countries that may not have Australia's best interests at heart. Todays [sic] friendly might be tomorrow's enemy". Others conveyed a general but solidified distrust with statements such as "allies are no longer trustworthy". Participants also noted the country's isolated geographic location, for example, "as an island continent, Australia is particularly vulnerable to supply disruption so reduces reliance on others (including being outbid by other customers)".

Participants make selections to optimize the technical functioning of the submarine.

For a large group of participants, the core desire for each design choice was a motivation to improve the likelihood that the automated system could complete its task successfully. This can be inferred from motivations for choosing to increase the batteries and sensors on the submarine. Participants who selected this combination frequently mentioned the importance of 'effectivity' and mission or task 'continuity' for the submarine. For example, one person who chose more sensors described their reason as "to avoid damage to the craft so they have a longer service life". Another participant who chose more batteries stated, "Better to have more battery capability for operational success" and "If it is capable of staying out for 3 days I think that this can save time

in the sense, if it only has one battery and needs to return to recharge you lose that day of investigating". Another theme for this group of participants was the importance of back-up support in sensors and batteries: "We should have a back-up sensor just in case it is needed" and (related to the need for the battery), "we need to get it [the submarine] to return to home base". This reason relates to the adjacent theme of retrieval, noted via words such as 'not getting lost', or 'returning'. Often, return was tied to references of completing a military objective (e.g., cost, not falling into an adversary's hands, or turnaround time for another deployment). For example, one participant justified the choice of batteries by arguing for the need "to ensure the submarine can get home", and another stated, "Increase the ability of recovery for reuse of the capability".

Participants who prioritized increasing batteries and sensors put high value and high trust in the technology and were willing to give up human control, although not unconditionally, as indicated by statements that allowed human control to be absent provided that the technology functions well. They use words such as "if" and provide adjectives related to the design or functioning, such as "correct" or "well". For example: "*As long as* the device is reliable and sensitive then autonomous functionality is fine, in fact probably preferable" [italics by authors]. Or another participant wrote, "*If* an autonomous submarine minesweeper can be *sufficiently good* at detecting and defusing/safely exploding underwater mines, this seems like something that does not need a lot of human control" [italics by authors].

Participants who do not trust the technology to various degrees.

Participants who somewhat increased human control (slider positions 2 and 3) did not completely trust or distrust the system. One participant, for instance, stated, "We can't *fully* trust artificial intelligence to make the best decisions" [italics by authors]. This group believed that some degree of human control was needed, as "sometimes technology needs to be override [sic] to correct technology error". A recurring theme for this group of participants was the prevention of mistakes or the malfunctioning of the system, as mentioned above. Different views exist on when this is needed. Some people's concerns over mistakes were in relation to external factors: "I am concerned about cybersecurity, and while I trust the machines, I would prefer to have more human control to mitigate any potential risks". Others mentioned human 'smartness' versus AI. For example, "An AI algorithm will never be smarter than a human". A small group of people thought we needed humans to prevent the systems from becoming too smart: "You gotta have human control or else the machines will rise against us r.e. The Terminator/Skynet".

Participants who maximized human control frequently made strongly voiced statements such as the following, "I 100% disagree with full automation" or "DONT LEAVE EVERYTHING TO AI AS ITS NOT TRUSTWORTHY IT THINKS FOR ITSELF ITS A FACT". They are largely motivated by concerns about autonomous systems that can "get out of control" or "take over the world" and consider human control necessary.

#### 4.4.2 Design task 2: the autonomous armed drone

Contrary to Design Task 1, participants did not immediately see how concrete design options (e.g., more battery capacity) related to design values (e.g., reliability) in the design task for the autonomous drone. The design task's primary interface included only concrete design options (Fig. 2), although participants could click the "i" button next to each design feature to see suggestions about the values that a feature could realize.

Participants made selections to prevent or mitigate errors.

Participants were worried about errors or mistakes by the system resulting in severe harm. They believed that human-assisted decision making could control autonomous systems and improve their trustworthiness. For instance, one participant mentioned, "I don't trust technology to decide what to drop the bomb on and would like human assistance to ensure innocent victims are not targeted". Another participant described a concern over incorrect targeting: "I feel like it would be safer to ensure the 'attack' that the drones are protecting us from isn't a party boat or something".

Such considerations for errors or mistakes were frequently mentioned by participants selecting 'video recordings of bombings'. Participants often mentioned that this can assist in fact-finding, verifying that the correct target has been hit, settling cases of disputes, or recording war crimes. These participants stated that it was "required in the event that there is a dispute as to the targets" or that it "provides evidence of (both) how well/how effectively the drones are working and that 'we' are not committing war crimes by using them". According to these participants, the video recordings could be used to prove that the right decision was made, whether by a human or machine, or in some cases to ensure accountability for the actions of the drone: "Important for accountability, especially potential war crimes. Allows for review of events and provides important data which could be used for other things".

Another group of participants chose 'video recordings' for systems improvement, which includes mitigating the risks of future errors. These participants provide pragmatic and functional reasons for selecting the video recording options: "Ability to review the accuracy and performance of the drone for future development and improvement in operations". They showed an acceptance that errors will occur and selected the design option of video recordings to minimize these errors. These participants seemed to choose video

recordings as an oversight mechanism supporting the usage of the drone. In contrast, other participants did not mention video recordings through this constructive lens. They instead saw error as something that disqualifies the use of the drone and stated that its actions should be recorded for the purpose of accountability.

Participants made selections with a concern for success and effectiveness.

Participants also frequently mentioned the importance of the drone getting the job done, so for them, effectiveness, mission success and endurability were key. Participants who selected 'redundancy in battery power' primarily argued that autonomous drones should be able to keep going. Deeper concerns about military effects predominantly underlie these arguments. One participant, for example, mentioned, "In a survival or invasion situation, the drone may need a longer life to ensure mission success". In addition, another described their reason as "so that a planned attack is not aborted because of battery problems". There are also participants who selected extra battery power for cost-effectiveness reasons, as this participant clearly illustrated: "If it's too expensive in the battlefield, it's important to be able to retrieve it for further use".

Participants who selected the 'swarm mode' (drones that are programmed to collaborate in swarms of multiple drones) also frequently explained this choice by referring to increased effectiveness, either in terms of the mission or in terms of the task. Regarding mission success, one participant, for instance, stated, "You can effectively overwhelm defensive efforts and successfully complete the mission". In terms of the task, another stated: "Greater effectiveness a chance of a few surviving to achieve task".

Participants made selections with an underlying concern for safety and security.

Participants made remarks about the need for back-ups to include a feature 'just in case', ensuring the system is available when needed. For example, when choosing an additional battery for the drone, one participant said, "It would be good to have a backup in case of failure". Another participant was concerned for the safety of troops and therefore chose an emergency communication system: "Essential if 'friendly fire' occurs".

Some participants mentioned specific Australian experiences of unsafe situations as reasons for adding emergency communication. For example, one stated, "I expect communications to be attacked first. Even during bushfires where I live, internet and phone were unavailable most of the time. So I chose this option". Others referred to the Australian geographic context with its vast areas of remoteness with no phone reception: "Communication in much of Australia is patchy at best so it's a no brainer".

Finally, some participants expressed security concerns in terms of the need for protection. For example, a participant who chose more battery power stated that "it needs to be able to operate for longer, what happens if the batteries die then we have russia invade our shores or the chinese, we need to be protected all the time".

#### 4.4.3 General observations from design task 1 and design task 2

Throughout the survey, we find evidence that participants considered the context in which they envisioned the technology to be used when making design choices. Frequently mentioned contextual features include Australia's geographic situation (i.e., vast land and water surfaces), location (i.e., far out relative to the rest of the world) and natural circumstances (flood and bushfire prone). In the case for Australia's self-sustainability, one participant described the following geographical feature: "As an island continent, Australia is particularly vulnerable to supply disruption so [this feature] reduces reliance on others". Reference to Australia's location was, for instance, made by a participant when choosing emergency communications: "I think having communication during an event would be extremely important. We live in a very remote area so any information to us would be crucial". Finally, participants also referred to Australia's natural disasters, such as this participant who chose the swarm mode: "To put drones into action in times of natural disaster as this is more prominent in our area."

We also found evidence that participants made trade-offs between design options in both design tasks. For example, in the case of the submarine, one participant chose to somewhat improve Australia's self-sustainability: "Tried to find a balance with other costs—but prefer australian made". Some participants traded human control for sensitivity and reliability, as the following quote illustrates: "I had better sensitivity and reliability, so it doesn't need personal input".

Finally, the qualitative analysis reveals that participants provide similar reasons for different design choices. Regarding the drone, the context in which it performs its task is mentioned as an argument for *increasing* human control, while in the case of the submarine, the context and task are mentioned as an argument for *decreasing* human control.

In addition, participants frequently mention that human control is necessary when the stakes are high. For example, one participant wrote, "Since these drones will be carrying lethal weapons (bombs) over areas where people live, you must have a human in the loop to make sure innocent people are not killed". In the case of the minesweeping submarine, on the other hand, participants frequently mentioned that human control was less important when the stakes are low. For example, one stated, "Not a great risk to people given the context of the problem (deep underwater) so not that big of an issue if the submarine makes the wrong decision".

### 4.5 Experience of the participants

Participants generally provided positive feedback in the survey. Participants especially appreciated that trade-offs could be made. For example, one participant stated, "I like the idea of design and compromise. Noone ever wants to discuss compromise". Another participant stated, "Way better than the census! […] I thought [the trade-off exercises] were about as well balanced as they could be. And honestly, they were probably better defined than some things that make it to AusTender [authors: Australia's government procurement information system;[…]]." Participants also appreciated the nuanced way in which they could share opinions. For instance, one stated, "I like how the limited options truly show black & white answers to an overwhelmingly grey problem." Finally, several participants mentioned the informative value of the survey, for example, one stated, "It gave me the ability to shape perceptions on a very important subject."

There were also participants who did not appreciate the survey method. Some participants believed that the subject matter was too technical to consult citizens about: "Do not ask every day average people such technical and calculated questions". Finally, some participants expressed distrust of the motivations behind the research, for example, "A completely ridiculous survey […]. Seriously, how much money is being wasted with confirmation bias trying to achieve an outcome that satisfies the paradigm that people don't want autonomous drones. […]".

## 5 Discussion

In this study, we set out to investigate which values people would like to see embedded in autonomous military systems. Using PVE, we consulted a large and representative group of Australians about which values matter in relation to two specific technologies: an autonomous minesweeping submarine and an autonomous drone that can drop bombs. Design options for both technologies were presented in an integral manner, and detailed, realistic and contextual information was provided about the meaning and consequences of design choices. In this way, we informed and actively engaged a broad public cohort with this complex topic. Moreover, by including a restriction (limited budget) in each design task, participants were forced to make trade-offs between design options and the values that these options realize. Our results indicate that when faced with a realistic choice, the 'general public' has diverse and nuanced stances on the question of human control over autonomous military systems, views that are more complex than the views expressed in the academic and international debates referenced in the introduction.

We find that a third of participants are opposed to autonomous military systems when asked about their opinion via a straightforward survey question (33%). However, when this group is asked to design concrete autonomous systems in a PVE, participants selected many different combinations of design features realizing varying degrees of human control. Several contextual factors, technology-specific concerns and certain values seemed to explain these different choices.

### 5.1 Participants were concerned about the reliability of autonomous systems. These people prioritized technology-centred solutions.

Participants who did not or who only incrementally increased the requirement for human control in the case of the submarine frequently opted to add more batteries and sensors instead. Our qualitative analysis reveals that these participants prioritized engineering solutions that directly increased the technical functioning of the system (rather than sociotechnical solutions, such as increasing human control or increasing Australia's self-sustainability) because they believed that this could prevent accidental harm to humans.

In the case of the drone, 'human-assisted decision making' was selected by a minority of participants (28%). The cluster analysis showed that participants with diametrically opposite views on the value of human control (realized through the design option 'human-assisted decision making') both prioritized redundancy in battery power and video recordings of bombings over other design features, although in a different order. The group of participants who did not select human control for the drone slightly prioritized battery power over video recordings, while the group of participants who all chose human control strongly prioritized video recordings over battery power. It can be inferred that people in favour of system autonomy (i.e., no or limited human control) are technology centred, in that they wanted to 'get the technology right'.

Another frequently recurring theme was the concern for retrieval. This was most prominent among participants who selected additional batteries, swarming functions, or emergency communications. Participants stated they were concerned about the effectiveness of the system and that they should be available to complete its task or mission. Such pragmatic reasons were also found in the motivations for selecting video recordings of bombings, which was surprising because the researchers had included this option to realize the value of 'accountability', but participants rarely mentioned this. Instead, participants often selected these features to settle disputes about fact finding or to improve the functioning of the system by learning from the recordings.

### 5.2 A minority distrusted autonomous systems and called for human-centred solutions

The cluster analysis shows that the cluster that univocally selected 'human-assisted decision making' for the drone has their second preference in a nontechnical solution as well (video recordings). In other words, this group prefers sociotechnical solutions rather than technical solutions (such as additional batteries or geographical restrictions) to mitigate the risk of error. Our qualitative analysis revealed that the group of participants who selected 'human-assisted decision making' did so because they did not trust autonomous systems or technology in general. This may explain why they were less inclined to trust technological or engineering fixes to mitigate the risk of errors and preferred sociotechnical design options.

### 5.3 Participants considered the context of use and risks to humans

We offered participants two different design tasks for autonomous systems that varied in the areas *where* they could be used and *how* they could be used. On the one hand, the submarine functions underwater with a low probability of encountering other people in normal circumstances. People may be accidently harmed when the submarine overlooks a diver in the area or when the submarine is hacked and taken over by adversaries. In the case of aerial drones, on the other hand, the presence of people is highly likely when the drone functions under normal circumstances because it is designed as a directly offensive weapon against people.

The qualitative data revealed that people considered these contexts by providing clues that they saw different risks in different contexts, allowing for less human control when the risks to humans were lower. This is an indication that participants did not necessarily distrust autonomous military systems as some surveys suggest but that their trust is context specific. Participants expressed their choices by increasing the reliability in the case of the submarine and verifiability and accountability in the case of the drone to mitigate risks to humans.

### 5.4 Australia's unique location drove concerns for self-sustainability

The qualitative data showed that many participants considered Australia's geographically isolated position in relation to the need to be self-sufficient. Participants expressed that Australia has to care take of itself, 'just in case', including multiple references to the need for 'back-ups'. These motivations reveal symptoms of what is colloquially called the 'tyranny of distance'.[2] It relates to the long geographic distance between Australia and the 'Mother Country' Great Britain, which has historically been a source of insecurity in discussions of Australia's economic and strategic future. Although the reality of the tyranny of distance has been contested, concerns for self-sustainability driven by Australia's unique geographic location seemed to play a large role in the explanation of why certain design features were particularly favoured by certain participants .

Sensitivity to a country's geographical situation and geopolitical history is important in the broader context of surveys about autonomous military systems. A survey in one country may elicit different value discussions around such systems than in other countries. Explanations in terms of varying cultural or political preferences may be too simplistic to explain the underlying concerns that inform the values people consider. There may be structural circumstances underlying people's concerns, such as the case of Australia's unique location on the world map.

### 5.5 Implications for engineering and policy

Our combined quantitative and qualitative analysis showed that there exist different subgroups in society with varying, nuanced concerns about autonomous systems. An implication for policymaking is that decision-makers must recognize that some solutions speak better to certain groups in society than others. Ideally, both technical and sociotechnical solutions are given full consideration when moving the discussion on autonomous systems forward.

Moreover, decision makers should recognize that the values and design requirements specified by experts do not necessarily capture the concerns of 'the public'. Our qualitative data show that people make their design choices with reference to different values from the values that the researchers tacitly 'assigned' to the design features based on consultations with experts. For example, the drone design task was designed so that 'video recordings of bombings' would suggest the values of 'accountability' and 'accuracy'. However, we found that many people who chose 'video recordings of bombings' argued that this would be good for verifiability, rather than accountability for bombings. For the case of 'human-assisted decision making', researchers posed human control as a value in and of itself (reminiscent of 'human autonomy'), but participants referred to fear of system mistakes or errors or to ensuring accountability when explaining their choice for human-assisted decision making. This shows the importance of qualitative data gathering when surveying participants' values on a topic.

[2] Blainy, Geoffrey, (Blainey 1996) The Tyranny of Distance: How Distance Shaped Australia's History. Melbourne: Sun Books.

The practical implications are that in communicating the risks and opportunities of design choices for autonomous systems, explanations are needed that account for people's (limited) understanding of military and engineering realities, which may otherwise be obvious to subject-matter experts.

### 5.6 Limitations

A limitation of this study is that we compare the design choices people make in two design tasks that varied along different dimensions, including context, setup and options. First, both tasks varied in context, as the first was about a nonlethal minesweeping submarine, whereas the second was about a lethal aerial drone. Second, both tasks consisted of different types of PVE: values could be increased in salience by implementing more of the design feature vs. values that could be highlighted only by selecting design options. Third, both tasks differed in terms of the values included. Due to these multiple differences, interpreting the difference between choice patterns (the LCCAs) in both design tasks is cumbersome, and we cannot clearly distinguish what causes certain values to be selected more frequently in one versus the other. Therefore, we primarily focus our comparative analysis of both cases on the qualitative analysis. Further research could investigate to what extent value trade-offs differ between different technologies when using exactly the same PVE set-up, or vice versa, how value trade-offs differ when using different PVE set-ups for the same technology. On a positive note, in the qualitative data we did not find many people that disagreed with how the researchers had translated values into design choices.

A further limitation of this research is that our survey design may not have suited some stakeholder groups well, such as people with physical disabilities (e.g., blind people). Attempts to discuss our PVE design with often overlooked stakeholder groups who may find the setup of the PVE particularly challenging for mental reasons (e.g., neurodiverse people) or due to a cultural mismatch (potentially people from indigenous groups) were unfortunately not successful.

Finally, external factors necessarily influenced this survey, and one such factor was the Russian invasion of Ukraine. This PVE was launched in March 2023, which was one year into the war. Reports and even videos of drones and other advanced technologies in use on the battlefield were in the public domain. Our qualitative data showed sporadic evidence that people were motivated by current events when making design choices. For example, one participant who chose an emergency communication system justified their reasons as "because denied and degraded is how the military starts planning. Ukraine has shown the importance of this against an enemy that does not care about war crimes like Russia, and China will be the same".

## 6 Conclusion and recommendations

Our PVE survey contributed to a nuanced discussion on values as they relate to autonomous military systems, with particular attention to the value of human control. We offered participants an integral and informed set of values that can be included in autonomous military systems. For the purpose of this article, we focused on different degrees to which human control could be exerted in autonomous systems design and how participants traded this value for other values.

The process of having to trade off design requirements revealed the participants' hierarchy of values. It also revealed that there are actually not two ends of the spectrum in discussions on autonomous military systems, but a large, nuanced majority and a very small group that express their motivations very strongly.

Our research has shown that value preferences are informed by underlying concerns directly related to a countries' geographic location and characteristics. Recommendations for further research therefore include doing a similar PVE in countries that have strategic or military partnerships with Australia, such as AUKUS, NATO and the Five Eyes partnership.

Value differences between countries that design and use autonomous military systems may only become clear once operationalized in joint exercises or international missions.

Furthermore, the PVE method had not been applied in Australia nor on this rather technical topic. We were unsure how this method would work, but we can conclude that we have conducted a PVE quite successful in this context, which is a scientific contribution.

By asking people to motivate their survey choices in writing, we captured value expressions in an alternative manner. Moreover, in some cases, it revealed some misalignment between the value interpretations of researchers and those of participants. This is an important insight for value-mining research as a method in the social sciences and for value-sensitive design research for ethics and design studies. Values may mean different things to different people in different contexts. It also shows that the task of translating values into design choices is not straightforward.

Our findings offer a unique empirical perspective on the value of human control in autonomous military systems and on other values and value trade-offs associated with the design of autonomous military systems. These can function as inputs for the value-sensitive design (cf. Friedman and Kahn (Friedman and Kahn 2003)) of autonomous military systems, as they lay out how different groups of people prioritize values, where values overlap and a conflict between values may never be overcome. These values

serve as a signpost for decision-makers, industry professionals and military organizations that take citizens' concerns seriously to move the discussions on autonomous military systems forward in a constructive manner.

**Acknowledgements** The research for this paper received funding from the Australian government through Trusted Autonomous Systems, a Defence Cooperative Research Centre funded through the Next Generation Technologies Fund.

**Curmudgeon corner** Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions.

**Data availability** The data that support the findings of this study are not publicly available. The data are, however, available for viewing from the authors upon reasonable request.

## Declarations

**Ethical approval** The Human Research Ethics Committee of UNSW approved our study protocol (HC220732).



## References

Australian government (2021). Defence FOI 187/20/21. Defence Instruction Administrative Policy.

Amoroso D, Tamburrini G (2020) Autonomous weapons systems and meaningful human control: ethical and legal issues. Curr Robot Rep 1:187–194

Arai K, Matsumoto M (2023) Public perceptions of autonomous lethal weapons systems. AI Eth. https://doi.org/10.1007/s43681-023-00282-9

Asaro P (2012) On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. Int Rev Red Cross 94:687–709

Blainey G (1966) The tyranny of distance: how distance shaped Australia's history. Sun Books, Melbourne

Blanchard A, Thomas C, Taddeo M (2024) Ethical governance of artificial intelligence for defence: normative tradeoffs for principle to practice guidance. AI Soc. https://doi.org/10.1007/s00146-024-01866-7

Bo M (2022) Are programmers in or 'out of' control? the individual criminal responsibility of programmers of autonomous weapons and self-driving cars. SSRN Scholarly Paper at https://papers.ssrn.com/abstract=4159762. Accessed 5 Feb 2024

Boshuijzen-van Burken C (2023) Value Sensitive design for autonomous systems in defence - a primer. J Eth Inform Technol. https://doi.org/10.1007/s10676-023-09687-w

Boshuijzen-van Burken C, Spruit S, Fillerup L, Mouter N (2023) Value sensitive design meets participatory value evaluation for autonomous systems in Defence. In: 2023 IEEE International symposium on ethics in engineering, science, and technology (ETHICS), pp 1–5. https://doi.org/10.1109/ETHICS57328.2023.10155025

Burton S et al (2020) Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artif Intell 279:103201

Carson R, Groves T (2007) Incentive and informational properties of preference questions. Environ Resour Econ 37:181–210

Cebulla A, Szpak Z, Howell C, Knight G, Hussain S (2023) Applying ethics to AI in the workplace: the design of a scorecard for Australian workplace health and safety. AI & Soc 38:919–935

Conboy C (2021) Opposition to killer robots remains strong — poll. *Stop Killer Robots* https://www.stopkillerrobots.org/news/poll-opposition-to-killer-robots-strong/. Accessed 5 Feb 2024

Deeney C (2019) Six in Ten (61%) Respondents across 26 countries oppose the use of lethal autonomous weapons systems. *Ipsos* https://www.ipsos.com/en-us/news-polls/human-rights-watch-six-in-ten-oppose-autonomous-weapons. Accessed 10 Jan 2024

Devitt K, Gan M, Scholz J & Bolia R (2021). A method for ethical ai in defence. https://apo.org.au/node/311150. Accessed 7 Dec 2023

Docherty B (2020) The need for and elements of a new treaty on fully autonomous weapons. Proceedings of rio seminar on autonomous weapons systems 20 february *2020* Alexandre de gusmão foundation, Rio de Janeiro

Ekelhof MAC (2018) Lifting the fog of targeting. Nav War Coll Rev 71:61–95

Ekelhof M (2019) Moving beyond semantics on autonomous weapons: meaningful human control in operation. Glob Policy 10:343–348

Friedman B. & Kahn P (2003) Human values, ethics and design. In the human-computer interaction handbook 1177–1201

Future of life institute (2015) Autonomous weapons: an open letter from AI & robotics researchers. Future of Life Institute

Golan E (2023) Step into the driver's seat: a participatory value evaluation of the public transport policy preferences of the Tel Aviv metropolitan area & israeli face validity analysis. Delft University of Technology, Delft

Gonzales Pecho H (2023) Climate change mitigation policy alternatives and citizens' preferences trade-offs: a participatory value evaluation in Peru. Delft University of Technology, Delft

GGE LAW (2019) Australia's system of control and applications for autonomous weapon systems. CCW/GGE.1/2019/WP.2/Rev.1

Hadlington L et al (2024) Public perceptions of the use of artificial intelligence in defence: a qualitative exploration. AI & Soc. https://doi.org/10.1007/s00146-024-01871-w

Horowitz MC (2016) Public opinion and the politics of the killer robots debate. Res Politics 3:205316801562718

Hössinger R, Peer S, Juschten M (2023) Give citizens a task: an innovative tool to compose policy bundles that reach the climate goal. Transp Res Part: Policy Pract 173:103694

Human rights watch (2016) Killer robots and the concept of meaningful human control—memorandum to convention on conventional weapons (CCW) delegates. https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control. Accessed 10 Dec 2023

Isaacson W (2023) Elon Musk. Simon and Schuster
Itten A, Mouter N (2022) When digital mass participation meets citizen deliberation: combining mini- and maxi-publics in climate policy-making. Sustainability 14:4656
Johnston RJ et al (2017) Contemporary guidance for stated preference studies. J Assoc Environ Resour Econ 4:319–405
Lillemäe E, Talves K, Wagner W (2023) Public perception of military AI in the context of techno-optimistic society. AI & Soc. https://doi.org/10.1007/s00146-023-01785-z
Mouter N et al (2022) Stepping into the shoes of the policy maker: results of a participatory value evaluation for the Dutch long term COVID-19 strategy. Soc Sci Med 314:115430
Mouter N, Koster P, Dekker T (2019) An Introduction to Participatory Value Evaluation. SSRN J. https://doi.org/10.2139/ssrn.3358814
Mouter N, Koster P, Dekker T (2021b) Participatory value evaluation for the evaluation of flood protection schemes. Water Resour Econ 36:100188
Mouter N, Koster P, Dekker T (2021c) Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. Transp Res Part : Policy Pract 144:54–73
Mouter N, Hernandez JI, Itten AV (2021a) Public participation in crisis policymaking. How 30,000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures. PLoS ONE 16(5):e0250614
NATO (2021). NATO review - an artificial intelligence strategy for NATO. NATO Review https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html. Accessed 10 Dec 2023
Owen, R. *et al.* (2013) A framework for responsible innovation. In: Owen, R., Bessant, J. & Heintz, M.(ed) Wiley, pp 27–50
Paraman P, Anamalah S (2023) Ethical artificial intelligence framework for a good AI society: principles, opportunities and perils. AI & Soc 38:595–611
Rosendorf O, Smetana M, Vranka M (2022) Autonomous weapons and ethical judgments: Experimental evidence on attitudes toward the military use of "killer robots." Peace and Confl: J Peace Psycholo. https://doi.org/10.1037/pac0000601
Rosert E, Sauer F (2021) How (not) to stop the killer robots: a comparative analysis of humanitarian disarmament campaign strategies. Contemp Secur Policy 42:4–29
Russell S (2023) AI weapons: Russia's war in Ukraine shows why the world must enact a ban. Nature 614:620–623
Sadek M et al (2024) Challenges of responsible AI in practice: scoping review and recommended actions. AI & Soc. https://doi.org/10.1007/s00146-024-01880-9
Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems a philosophical account. Robot Front. https://doi.org/10.3389/frobt.2018.00015
Steen M, van Diggelen J, Timan T, van der Stap N (2023) Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives. AI Ethics 3:281–293
Taddeo M, Blanchard A (2022) A Comparative analysis of the definitions of autonomous weapons systems. Sci Eng Ethics 28:37
Taebi B, Correljé A, Cuppen E, Dignum M, Pesch U (2014) Responsible innovation as an endorsement of public values: the need for interdisciplinary research. J Responsib Innovation 1:118–124
UK DoD (2022) Ambitious safe responsible: our approach to the delivery of AI-enabled capability in Defence
US DOD (2023) Dod directive 3000.09 autonomy in weapon systems. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.PDF?ver=e0YrG458bVDl3-oyAOJjOw%3d%3d
United Nations (2023) Note to correspondents: joint call by the United Nations Secretary General and the president of the International Committee of the Red Cross for States to establish new prohibitions and restrictions on autonomous weapon systems | united nations secretary-general. https://www.un.org/sg/en/content/sg/note-correspondents/2023-10-05/note-correspondents-joint-call-the-united-nations-secretary-general-and-the-president-of-the-international-committee-of-the-red-cross-for-states-establish-new. Accessed 5 Feb 2024
Verdiesen I (2017) How do we ensure that we remain in control of our autonomous weapons? AI Matters 3:47–55
Verdiesen I, Dignum V (2022) Value elicitation on a scenario of autonomous weapon system deployment: a qualitative study based on the value deliberation process. AI Ethics. https://doi.org/10.1007/s43681-022-00211-2
Verdiesen I, de Sio FS, Dignum V (2019) Moral values related to autonomous weapon systems: an empirical survey that reveals common ground for the ethical debate. IEEE Technol Soc Mag 38:34–44
Vermunt JK, Magidson J (2013) Technical guide for latent GOLD 5.0: basic, advanced, and syntax. Statistical Innovations Inc, Belmont MA
Young KL, Carpenter C (2018) Does science fiction affect political fact? yes and no: a survey experiment on "killer robots." Int Stud Quart 62:562–576
Zhang B et al (2021) Ethics and governance of artificial: intelligence evidence from a survey of machine learning researchers. Jair. https://doi.org/10.1613/jair.1.12895