



# Variable Autonomy through Responsible Robotics: Design Guidelines and Research Agenda

TYLER REINMUND, PERICLE SALVINI, LARS KUNZE, and MARINA JIROTKA, University of Oxford, UK  
ALAN F. T. WINFIELD, UWE Bristol, UK

Physically embodied artificial agents, or robots, are being incorporated into various practical and social contexts, from self-driving cars for personal transportation to assistive robotics in social care. To enable these systems to better perform under changing conditions, designers have proposed to endow robots with varying degrees of autonomous capabilities and the capacity to move between them—an approach known as variable autonomy. Researchers are beginning to understand how robots with fixed autonomous capabilities influence a person's sense of autonomy, social relations, and, as a result, notions of responsibility; however, addressing these topics in scenarios where robot autonomy dynamically changes is underexplored. To establish a research agenda for variable autonomy that emphasises the responsible design and use of robotics, we conduct a developmental review. Based on a sample of 42 papers, we provide a synthesised definition of variable autonomy to connect currently disjointed research efforts, detail research approaches in variable autonomy to strengthen the empirical basis for subsequent work, characterise the dimensions of variable autonomy, and present design guidelines for variable autonomy research based on responsible robotics.

CCS Concepts: • **Computer systems organization** → **Robotics**; • **Human-centered computing** → **Interaction design**;

Additional Key Words and Phrases: Variable autonomy, responsible innovation, responsible robotics, literature review

## ACM Reference format:

Tyler Reinmund, Pericle Salvini, Lars Kunze, Marina Jirotko, and Alan F. T. Winfield. 2024. Variable Autonomy through Responsible Robotics: Design Guidelines and Research Agenda. *ACM Trans. Hum.-Robot Interact.* 13, 1, Article 7 (January 2024), 36 pages.  
<https://doi.org/10.1145/3636432>

## 1 INTRODUCTION

Robots are being incorporated into various practical and social contexts, from self-driving cars for personal transportation to assistive robotics in social care. There is an emerging understanding

This work was conducted within the project RoboTIPS: Developing Responsible Robots for the Digital Economy supported by EPSRC grant ref EP/S005099/1.

Authors' addresses: T. Reinmund, P. Salvini, and M. Jirotko, Department of Computer Science, University of Oxford, 39a St Giles., Oxford UK OX1 3LW; e-mails: {tyler.reinmund, pericle.salvini, marina.jirotko}@cs.ox.ac.uk; L. Kunze, Oxford Robotics Institute, University of Oxford, 23 Banbury Rd, Oxford UK, OX2 6NN; e-mail: lars@robots.ox.ac.uk; A. F. T. Winfield, Bristol Robotics Laboratory, UWE Bristol, Bristol UK, BS16 1QY; e-mail: alan.winfield@bri.ac.uk.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2024 Copyright held by the owner/author(s).  
2573-9522/2024/01-ART7  
<https://doi.org/10.1145/3636432>

of how robots with fixed autonomy influence a person's sense of autonomy, social relations, and, as a result, notions of responsibility [48, 72, 120, 123]. For example, some scholars have suggested that social robots in care homes can increase residents' feelings of autonomy by decreasing their dependence on staff [99] or helping them stay connected with friends and family through video [112]; meanwhile, others offer opposing critiques, claiming that social robots provide illusory and inauthentic social relations that may emotionally manipulate care home residents [44, 129].

But what happens when these robots are imbued with the potential to operate along a continuum of autonomous capabilities? We refer to this approach to robotics as *variable autonomy* [25]. Past work has shown that dynamically changing between LoAs in complex settings can improve a robot's performance [87]. For example, a robot for disaster response may need to operate in environments with limited network conditions (e.g., [79]). When communication channels are operating properly, a remote human operator can directly control the robot; in this situation, the robot is in a teleoperated mode and consequently has lower levels of autonomous capabilities. Yet, when there are instances of low connectivity between the teleoperator and robot, the robot may have to transition to a state of greater autonomous capabilities to perform its rescue task without direct control from its human operator. Another example when variable autonomy may be required comes from the domain of assistive robotics. Consider a care robot that supports medication management for older adults. For some adults, the robot may only need to provide auditory reminders at set times throughout the day. However, for those who suffer from dementia, the robot may have to perform a wider range of tasks at higher LoAs, such as physically moving throughout a house. Apart from having to change autonomous capabilities in accordance with individual differences across a group of users, we can also imagine how the care robot's autonomy may have to adapt to the same individual user's condition if it were to deteriorate over time. These examples show how variable autonomy implementations lead to situations in which control authority over a robot shifts between a human and artificial agent in response to some set of conditions.

Recently, others have proposed variable autonomy as a means to operationalise responsibility in the design of autonomous systems (see [82]). Instead, we take the opposite direction: how can we ensure that robots with variable autonomy are designed and developed in a responsible manner? The preceding scenario of the care robot for medication management highlights the potential risks of introducing variable autonomy into sensitive environments: too great or too little autonomy under certain conditions may result in various harms, such as missed medication or a person losing their sense of independence. The capability to alter a robot's autonomous capabilities during interaction accentuates questions pertinent to responsible robotics, such as the following. Under what environmental and social circumstances is variable autonomy appropriate? Who may be harmed and how? Who should be held accountable if control over a robot's capabilities may alter unexpectedly? To date, few, if any, studies have addressed the connection between responsibility and variable autonomy (see the work of Small et al. [116] for one such study), and none, as far as we are aware, have approached variable autonomy through the lens of responsible robotics.

Therefore, our objective in this work is to construct a research agenda for variable autonomy based on responsible robotics. To do so, we must first establish a coherent representation of variable autonomy research. In its present state, this field lacks cohesive terminology, leading to disjoint research efforts; a detailed description of the field's research approaches, making it difficult for scholars to adopt similar designs, employ consistent and validated measures, and identify empirical gaps; and a clear discussion of variable autonomy design guidelines that can serve as a heuristic for engineers and researchers. From these gaps and in pursuit of our objective, we address the following research questions:

*RQ1* : How is variable autonomy defined in the literature?

*RQ2* : How is research into variable autonomy conducted?

*RQ3* : How is variable autonomy implemented?

In answering these questions, we develop a novel model to the study and design of variable autonomy robotics that builds on prior empirical and conceptual research. This research model will be articulated through clear, consistent terminology, and guided by an in-depth understanding of past empirical approaches. Given these aims, we follow the “developmental review” method as described by Templier and Paré [126]. A developmental review is a structured literature review method from the field of information systems that is useful for developing novel conceptualisations, frameworks, and approaches from previous bodies of research. As our review, we survey 42 recent contributions to variable autonomy in robotics published in high-quality and high-impact venues; we expand upon our method in Section 3.

Based on our review, we make four contributions:

- (1) We present a synthesised definition of variable autonomy in robotics to provide a shared language for researchers, thereby linking together currently disjointed research efforts.
- (2) We detail research approaches employed in the literature; our focus is on the research designs, sites, and evaluative measures. Explicating these facets supports the development of rigorous experimental protocols in variable autonomy research, highlights the need to refine quantitative measures, and reveals the sheer absence of qualitative evidence surrounding people’s experiences with variable autonomy robots.
- (3) We distil previous characterisations of variable autonomy; this creates a heuristic for designers when considering requirements for variable autonomy robotics.
- (4) We present 11 design guidelines that will help researchers approach variable autonomy through a lens of responsible robotics. These guidelines cover both the product and process of variable autonomy research, and encourage an anticipatory and reflective approach that engages with a range of stakeholders.

This article is organized as follows. Section 2 presents a brief background on three relevant concepts—autonomy, variable autonomy, and responsible robotics—and a summary of related reviews. Then, in Section 3 we elaborate our developmental review method. We provide a brief description of our dataset in Section 4 before presenting our results in Section 5. Next, we present the 11 design guidelines for variable autonomy based on responsible robotics in Section 6. In Section 7, we discuss areas for future research at the intersection of variable autonomy and responsible robotics. Finally, we conclude by summarising our key findings.

## 2 BACKGROUND

In this section, we provide a background on three concepts relevant to this review. First, we specify our conception of autonomy as it relates to robotics. Second, we present a historical overview on variable autonomy and its motivating questions. Finally, we introduce the interdisciplinary topic of responsible robotics and its roots in responsible innovation. These three concepts are briefly defined in Table 1.

### 2.1 Autonomy

Autonomy is a contested concept, subject to centuries of moral and philosophical debate. It conjures notions of free will, self-governance, and independence. Consolidating a balanced perspective on what is meant by the term is therefore a knotty endeavour. Rather than engage in any philosophical pretensions, we instead describe autonomy as it is conceived in the robotics literature.

An important first step is to note the subtle distinction between autonomy and its etymological descendent: automation. Autonomy implies the ability to perceive, decide, and act independent of

Table 1. Brief Definition of Concepts Addressed in This Article

Concept	Definition
Autonomy	“The extent to which a robot can sense its environment, plan based on that environment, and act upon that environment with the intent of reaching some task-specific goal (either given to or created by the robot) without external control” [9, p. 77].
Variable autonomy	An interaction strategy between human and robot agents in which the robot’s level of autonomy varies during operation in response to changes in context [Source: authors’ analysis].
Responsible robotics	“Responsible robotics is the application of Responsible Innovation in the design, manufacture, operation, repair, and end-of-life recycling of robotics, that seeks the most benefit to society and the least harm to the environment” [141, p. 173].

an external force [130]; automation meanwhile refers to “the execution by a machine, usually a computer, of a function previously carried out by a human” [96, p. 931]. Autonomy is thus a more demanding concept when ascribed to a robot, requiring the capacity to both deliberate and act upon the world. A framework of autonomy in **Human-Robot Interaction (HRI)** by Beer et al. [9] provides the following definition: “The extent to which a robot can *sense* its environment, *plan* based on that environment, and *act* upon that environment with the intent of reaching some task-specific goal (either given to or created by the robot) without external control” [italic emphasis added] (p. 77).

As per Beer et al. [9], any task is composed of three “primitives”: sense, plan, act; a robot’s ability to perform each of these facets independently determines how autonomous it is said to be. Since the degree to which a robot executes each task primitive autonomously can vary, researchers have conceptualised robot autonomy in a hierarchical structure of potential control modes.

Taxonomies for **Levels of Autonomy (LoA)** have a long history within the automation and HRI literature. We do not attempt to delineate them all here, but focus on a few key contributions that help explicate the concept of LoAs. Those interested in greater detail can refer to reviews by Vagia et al. [130] and Beer et al. [9].

One of the earliest comes from research on automation by Sheridan and Verplank [114]. Published in 1978, the authors survey the potential of teleoperated and supervisory control systems: teleoperation means, intuitively, that a vehicle is controlled remotely by a human operator, whereas supervisory control includes vehicles that can operate automatically for periods of time with intermittent intervention by a remote operator. These control modes represent 2 of 10 potential levels; as one moves up the hierarchy, the extent to which human intervention is necessary decreases.

Building on this work over two decades later, Parasuraman et al. [98] expanded the framework to include both *types* and *levels* of automation. As before, automation varies across a continuum from manual performance to full automation. However, in this framework, the authors specified the classes of functions to which automation can be applied: information acquisition, information analysis, decision and action selection, and action implementation [98]. Automation is not all-or-nothing and can be applied to varying degrees to certain types of functions.

Alongside the proliferation of such taxonomies in the automation literature, researchers in HRI have articulated their vision of robot autonomy, taking into consideration the idiosyncrasies of robotics technology such as physical embodiment and social situatedness [9]. From the perspective of military applications, Huang et al. [54] created a framework to describe the LoAs along three

dimensions: the complexity of the mission, the difficulty of the environment, and the degree to which humans interface with the robot. Each axis contains a series of metrics which are used to calculate the robot's level of autonomy. In situations characterised by low mission complexity, simple environments, and a high degree of human interaction, the robot's autonomy is considered to be low, and the more independently the robot can sense, plan, and act during complex missions in difficult environments, the more higher LoAs are needed [54].

The most recent work by Beer et al. [9] sets out a framework that specifies in detail 10 different levels of robot autonomy. Across each level, the framework states the roles performed by both the human and robot, as they relate to the primitives of sense, plan, and act. For example, in a level titled *batch processing*, “[b]oth the human and robot monitor and sense the environment. The human, however, determines the goals and plans of the task. The robot then implements the task” [9, p. 87]. As one moves along the continuum from manual to full autonomy, the number of functions allocated to the robot increases.

Apart from its adoption in other academic disciplines, the level of autonomy concept has been profoundly influential in shaping international standards. For example, the SAE J3016 standard for “Levels of Driving Automation” depicts degrees of automation for vehicles [56], ranging from Level 0, in which the human manually operates all driver support features, to Level 5, where the automation drives the vehicle under any condition.

While influential, the LoA concept has been criticised by numerous authors. These critiques commonly take issue with the implied tradeoff between human and autonomous control, albeit through slightly different formulations. For example, Bradshaw et al. [16] implore that increases in a system's autonomy do not necessarily entail a concomitant decrease in the need for human control. Ironically [5], the introduction of an autonomous system tends to create new kinds of cognitively demanding work for human operators to perform [16]. Relatedly, Endsley [41] points to the *automation conundrum*: “The more automation is added to a system, and the more reliable and robust that automation is, the less likely that human operators overseeing the automation will be aware of critical information and able to take over manual control when needed” (p. 8). Building the line of critique levied against the LoA taxonomy, Shneiderman [115] proposes a two-dimensional framework in which high levels of human control and autonomous capabilities are simultaneously achievable.

In summary, frameworks for LoAs originate in the field of automation research and have been influential in numerous areas. Those involved in HRI have adapted these taxonomies to fit the nuances of robotics technology. The continuum of autonomy supposes that as the degree to which a robot can sense, plan, and act in its environment increases, the level of human involvement subsides. Despite its adoption in technical standards and much academic writing, the uni-dimensional LoA concept is heavily criticised.

## 2.2 Variable Autonomy

A central assumption of these frameworks is that LoAs are fixed at the design stage—what Parasuraman et al. [97] termed *static automation*. The resultant rigidity of these robots comes with various challenges, such as ensuring operators can intervene during automation failures [36, 37, 97] and enabling human-robot teams to adapt to changing and complex environments [104]. To accommodate the challenges presented by fixed LoAs, substantial research has been directed towards approaches that aim to dynamically shift between modes of autonomous control [37, 90]—which we call *variable autonomy*. As early as the 1970s, variable autonomy has appealed to roboticists; it promised flexibility amid dynamic and hostile environments, reduced workload for human operators, and the ability to exploit the complementary skill sets of humans and robots [47, 49, 66, 109]. The past four decades have seen a number of research groups investigate variable autonomy under

many different labels, such as traded control [65], adaptive autonomy [35], adjustable autonomy [18, 37], sliding autonomy [20, 36], and dynamic autonomy [21]. The different uses of these terms are discussed in further detail in Section 5.1.

Despite their shared concern for the limitations posed by fixed LoAs in robots, these similar concepts are loosely defined and inconsistently compared and contrasted: some authors provide similar definitions for different terms, some create subtle distinctions between them, and others offer no definition at all. This semantic ambiguity complicates attempts to formally characterise variable autonomy and unnecessarily separates related research efforts. In this section, we provide a historical background on the concept of variable autonomy in robotics, point to seminal work in the field and its motivating problems, and outline limitations in current taxonomies of variable autonomy to emphasise the need for a robust definition and characterisation.

One of the earliest formulations of the notion that a robot can possess multiple LoAs comes from the previously discussed report by Sheridan and Verplank [114], who distinguished between two types of control, which they term as *shared* and *traded*. As the authors wrote: “Here, to share control means that both human and computer are active at the same time. To trade control means that at one time the computer is active, at another the human is” [114, Section 6.1]. Shared control, as defined in a recent review, is a control mode in which “human(s) and robot(s) are interacting congruently in a perception-action cycle to perform a dynamic task that either the human or the robot could execute individually under ideal circumstances” [1, p. 511]. As such, a robot with shared control is not necessarily one with variable autonomy; it is a form of collaboration, typically described as a specific LoA [9], that aims to achieve a given task through complementary human-robot capabilities. Meanwhile, the distinction by Sheridan and Verplank [114] implies that traded control is a type of variable autonomy in which control of a robot is at any time in one of two discrete states: fully autonomous or remotely controlled [65].

Beginning in the late 1990s, the concept of variable autonomy and its variants took hold in robotics research. A 1999 symposium titled *Agents with Adjustable Autonomy* hosted by the AAAI brought together early contributors and offered an initial definition. According to the symposium co-chairs, “adjustable autonomy means dynamically adjusting the level of autonomy of an agent depending on the situation” [90]; the authors go further and state that adjustments in autonomy can be initiated by either human or autonomous agents. Some of the earliest studies on variable autonomy addressed its applications in diverse contexts such as space missions [17, 37] and urban search and rescue [21]; investigated the problem of coordinating control in human-robot teams [109, 138]; evaluated how changes in LoA affect task performance, situation awareness, workload, and acceptance [49, 77]; and designed user interfaces for controlling the autonomy levels of multiple robots [47], moving across a continuum of LoAs [36], and delegating planning tasks to autonomous agents [84]. As this research progressed, it began to revolve around several central problems: who initiates changes in autonomy, for what reason, and when [82, 87, 104].

### 2.3 Responsible Robotics

To achieve our objective of constructing a research agenda for variable autonomy based on responsible robotics, we must first define what responsible robotics is. In the past few years, numerous authors have attempted to provide a description that captures the dynamic and diverse landscape of research on the social and ethical issues associated with robotics. In a special issue of *Frontiers in Robotics and AI*, Brandão et al. [19] outline the aims of responsible robotics; as per these authors, the field “should focus both on *identifying* social and ethical issues, and on *designing* methods to account for (and alleviate) such issues” [emphasis in original]. Meanwhile, another special issue edited by van Wynsberghe and Sharkey [136] defines responsible robotics as “the responsible research and innovation of robot development processes as well as the resulting products of such

processes.” Along similar lines, Winfield et al. [141] provide the following definition for responsible robotics: “Responsible robotics is the application of Responsible Innovation in the design, manufacture, operation, repair, and end-of-life recycling of robotics, that seeks the most benefit to society and the least harm to the environment.”

From these three articulations, we see that responsible robotics is an instantiation of **Responsible Innovation (RI)** within the domain of robotics. RI, then, is described as an approach that aims to align the products and processes of research and innovation with societal values and expectations (see [106, 122]). Numerous authors have contributed to the conceptual foundations of RI over the past decade; therefore, we draw on this extensive corpus to sharpen the concept of responsible robotics. In doing so, we clarify terms in the preceding definitions that have multiple, and oftentimes opaque, meanings in the literature: responsibility, innovation, approach, and societal values.

In their synthesis of moral responsibility and responsible innovation, Van de Poel and Sand [133] distinguish between two interpretations of responsibility. The first, backward-looking responsibility, focuses on assessing a past sequence of events to attribute blame or praise for some outcome. It requires “the ability and willingness to account for one’s actions and to justify them to others” [133]. And the second, forward-looking responsibility, entails an obligation to ensure that some future state comes about. This interpretation of responsibility implies anticipation of innovation outcomes on the part of those involved in the innovation process. Given the inherently uncertain nature of innovation and the unpredictability of its outcomes, attributing forward-looking responsibility for the breadth of an innovation’s social, environmental, and ethical effects is challenging to adopt in practice [12].

The term *innovation* itself likewise has many faces in the RI literature. Van den Hoven [134] offers one such definition: “Innovation is an activity or process which may lead to previously unknown designs pertaining either to the physical world (e.g., designs of buildings and infrastructure), the conceptual world (e.g., conceptual frameworks, mathematics, logic, theory, software), the institutional world (social and legal institutions, procedures and organization) or combinations of these, which—when implemented—expand the set of relevant feasible options for action, either physical or cognitive” (p. 80). From this articulation, at least two interpretations of innovation are apparent: innovation as both a product and a process. The latter represents the act of innovating, whereas the former is the result. Other scholars have extended that definition to include both the purpose (the reasons motivating innovators [122]) and people (those involved in innovation activities [58]).

Within the past decade, several academic and policy organisations have formulated multiple RI approaches. Two of the most prominent are those presented by von Schomberg [106] and Stilgoe et al. [122]. From the world of policy, the EPSRC, the UK’s main funding body for engineering and physical sciences research, has assimilated the work of Stilgoe et al. into its “AREA” framework for RI [93], constituted by four dimensions: anticipate, reflect, engage, act [42]. For clarity in writing, we present the dimensions here as though they are discrete; in practice, they overlap and build on one another.

First, anticipation refers to structured processes to identify and evaluate potential future scenarios and their associated impacts: both intended and unintended, positive and negative [122]. As mentioned previously, innovation is rife with uncertainty; therefore, the goal is not accurate prediction, but anticipation of plausible and desirable futures towards which we guide innovation [74]. Second, reflection involves questioning underlying motivations, purposes, and assumptions, and understanding the boundaries of knowledge [122]. Third, engagement is the inclusion of diverse stakeholder groups throughout the innovation process, enabling deliberation and debate during anticipation and reflection. Despite the consensus in the literature that stakeholder engagement

is essential for responsible innovation [74, 132], questions remain on how to engage stakeholders with vastly distinct, and potentially incompatible, worldviews [12] and enable meaningful engagement [108]. Finally, acting is about using the insights gained from the three prior dimensions to guide innovation along desired trajectories.

Innovators are then tasked with shepherding innovations according to the values of various societal actors. But what exactly are values, and how are innovators meant to identify them? Value sensitive design, an approach that seeks to engage with human values during design processes, offers some help; as per Friedman and Hendry [46], values are “what is important to people in their lives, with a focus on ethics and morality” (p. 24). Yet, as Boenink and Kudina [13] argue in their critique of values in RI, values are not “pre-given stable entities, ready made for reflection” (p. 452). The meaning of a given value varies: from person-to-person, place-to-place, and time-to-time. The dynamism of values has implications for innovators’ strategies to identify them. One method is to appeal to *a priori* defined lists of ethical principles. Such lists offer a helpful starting point and heuristic for dealing with values in design [46]; however, a strict reliance on so-called “universal” values neglects those that are culturally contingent [14]. Therefore, other authors advocate for an empirically led approach to the identification of values, engaging with people in their place and practice to understand what it is they find important [13]. A common critique against this line of thinking is that it falls victim to the naturalistic fallacy—that is, it assumes that the things people value are those they *should* value [14]. Our own perspective sees merits in both strategies. As mentioned, pre-defined ethical guidelines provide a helpful basis for agreed-upon values. Yet, we also acknowledge that they should not be used too rigidly; it is crucial to consider the actual experiences of those impacted by a technology. Therefore, we draw from both strategies, noting how ethical guidelines can inform our understanding of values, but they must be complemented with an empirical investigation of those involved.

Responsible robotics applies elements of RI to the robotics innovation lifecycle to reach societal and environmental objectives. Responsibility for events that have yet to occur and those that have already come about is essential; the former depends on anticipatory practices, and the latter on transparency into past events and a causal understanding that links actions and outcomes. Innovation in robotics refers to its dimensions of process, product, purpose, and people: the how, what, why, and who of innovation. And following an RI approach emphasises anticipation of potential pathways, reflection on motivations and assumptions, inclusive deliberation with impacted stakeholders, and responsiveness to the insights brought up through this process. We ground our approach to interpreting societal values in ethical guidelines for robotics, most of which agree that these systems should not harm individuals or the environment, promote human rights and well-being, maintain transparency, and ensure that human designers and operators remain responsible and accountable [140]. International standards such as BS 8611:2016 “Guide to the Ethical Design and Application of Robots and Robotic Systems” [55], the IEEE 7000-2021 standard for “Model Process for Addressing Ethical Concerns during System Design” [118], and the IEEE 7001-2021 standard for “Transparency of Autonomous Systems” [117] have been built on top of these shared principles. However, we equally emphasise that any study must include opportunities to reflect on stakeholder values as they exist in their time and place.

## 2.4 Past Reviews

Researchers have conducted reviews that address similar topics to those covered in this study, as shown in Table 2. In an early paper, Bradshaw et al. [15] conducted a narrative review to distinguish the dimensions along which autonomy can be adjusted. Per Bradshaw et al. [15], autonomy includes both actions that one is capable of performing and those that one is allowed to perform; as such, a robot’s autonomy can be adjusted according to what it is allowed to do, what it is required

Table 2. Summary of Related Work

Reference	Period	Aspect		
		Robotics	Responsible Robotics	Methodology
Bradshaw et al. [15]	1996–2004	◐	○	○
Mostafa et al. [87]	2003–2015	◐	○	◐
O'Neill et al. [94]	1999–2019	○	○	●
Selvaggio et al. [110]	1989–2021	●	○	○
This study	2010–2021	●	●	●

○ indicates that a review does not focus on a given aspect, ◐ indicates that a review partially focuses on a given aspect, and ● indicates that a review directly focuses on a given aspect.

to do, what others think it could plausibly do, and what it is able to do. This initial taxonomy provided a helpful conceptualisation of the elements of autonomy that can be altered, but it did not offer any insight into other dimensions of variable autonomy, such as who adjusts and why. More recently, Mostafa et al. [87] performed a systematic literature review to map the extent of research on variable autonomy for multi-agent systems. Their review specifies six design requirements: how autonomy is defined, measures to evaluate autonomy, available autonomy modes, which agent controls changes in autonomy states, patterns of human-agent interaction, and techniques to evaluate autonomy adjustments. Selvaggio et al. [110] provided a brief narrative review on shared control and shared autonomy in robotics. In this review, the authors' definitions of shared control and shared autonomy resemble the distinction between adjustable and adaptable autonomy, respectively, as detailed in Section 5.1. Finally, O'Neill et al. [94] conducted a critical review on teamwork in human-autonomy teams. Importantly, their work excluded research on robotics because of the idiosyncrasies that arise from physical embodiment.

This review differs from existing work across four aspects:

- (1) *Period*: This review focuses on recent developments in variable autonomy for robotics, extending 6 years beyond the review by Mostafa et al. [87]. Although the review by Selvaggio et al. [110] aims to cover recent research, the authors did not intend to conduct a comprehensive survey and therefore did not include details on the time frame of papers included in their review.
- (2) *Robotics*: Whereas others have included both embodied and non-embodied artificial agents in their reviews [15, 87], we focus specifically on robotics. A robot's physically embodied nature allows them to move throughout and act upon an environment, as well as engage with people, in ways that traditional automation cannot [9]. Therefore, focusing on robotics specifically enables us to engage with the technology's idiosyncrasies.
- (3) *Responsible robotics*: The objective of this review is to establish a research agenda for variable autonomy that is based on responsible robotics. In contrast, the objectives of related work have been to construct general frameworks [15, 87] or synthesise existing research [94, 110]. As far as we are aware, this is the first study to focus on how variable autonomy can be approached through a responsible robotics lens.

- (4) *Methodology*: This study reviews the research designs, empirical sites, and evaluation measures employed in variable autonomy robotics research. In this sense, this review is similar to that of O'Neill et al. [94]; yet, as mentioned, their review explicitly excluded research on robotics. Meanwhile, Mostafa et al. [87] only briefly touched on the methodology of variable autonomy for robotics, stating that “[m]ost of the adjustable autonomy research results are obtained based on simulation programs . . . [and] hence, the results might lack valid testing” (p. 181). We strengthen their claim by providing evidence that the results of variable autonomy research may lack ecological validity given that most studies have been conducted in artificial settings, such as simulations or contrived physical environments.

### 3 METHOD

Because of the unresolved conceptual and operational ambiguities surrounding variable autonomy, and our objective of specifying an approach to variable autonomy that is based on responsible robotics, we employ a “developmental review” as proposed by Templier and Paré [126].

#### 3.1 Search Strategy

To account for the diverse terminology in variable autonomy, we employ three data collection strategies: database, backward, and forward searches. First, we query four databases: ACM Digital Library Full-Text Collection, IEEE Xplore, Elsevier Scopus (Scopus), and Clarivate **Web of Science (WoS)**. The first two databases provide comprehensive coverage of papers published in ACM and IEEE conferences, prominent associations for computing and technology research. The latter, Scopus and WoS, likewise are known to have extensive and high-quality coverage of journals and conferences [86]. We construct keyword searches for each database based on terms identified in previous reviews [87], consultations with researchers in variable autonomy, and informal database searches. The resultant keyword queries are shown in Appendix A. Searches were performed in January 2022. To focus on recent developments in the field, we restrict our search from 2010 to 2021. Additionally, we only include results published in journal articles or conference proceedings, and written in English. This strategy yields a total of 294 papers.

Additionally, we conduct a backward search by reviewing the reference lists of previous reviews and papers recommended by colleagues to identify further references. In parallel, we record seminal early works in variable autonomy based on recurring citations in papers identified through the database search; these include the following works: [3, 6, 17, 18, 21, 37, 47, 59, 60, 76, 77, 84, 90, 103, 104, 109, 138]. Next, we conduct forward searching in the Scopus database—retrieving papers that cite the previously stated seminal works or the review by Mostafa et al. [87]. Together, backward and forward sampling result in an additional 438 papers.

#### 3.2 Data Selection

Overall, our three search strategies lead to 732 papers. We then employ a multi-stage selection approach to identify relevant and representative papers. First, we remove any duplicate entries. Then, we review the titles and abstracts according to the following inclusion criteria:

- (1) Primary research: conceptual or empirical.
- (2) Full text is available.
- (3) Discusses an architecture or implementation of variable autonomy.
- (4) Discusses variable autonomy in relation to robotics (i.e., physically embodied artificial agents).

After this initial inclusion review, we are left with 154 papers. Given that we do not intend to provide an exhaustive review of the literature, we prioritise studies based on their publication

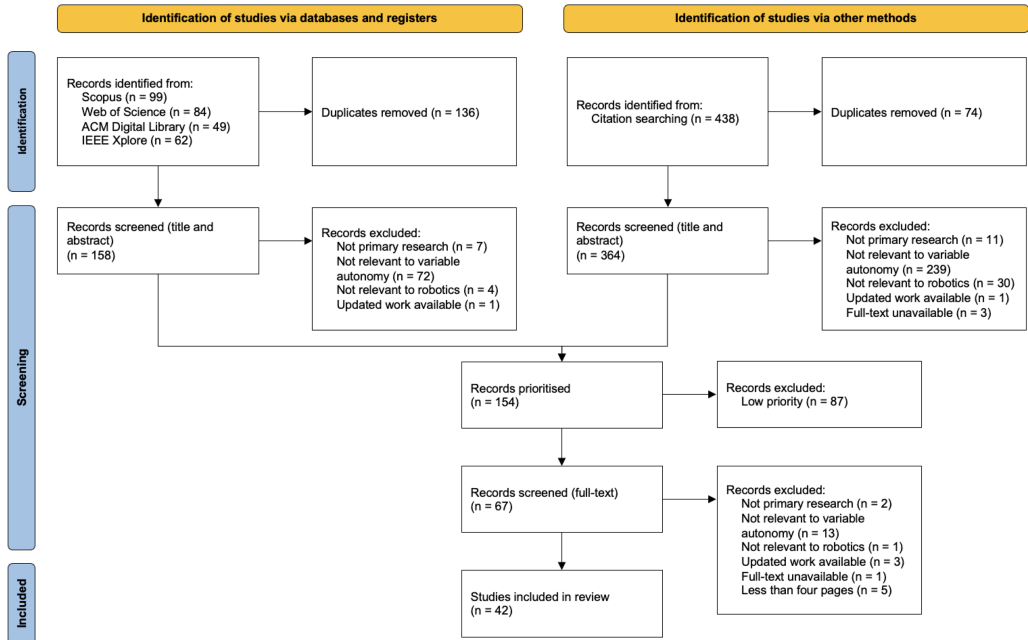


Fig. 1. PRISMA-style flowchart depicting the sampling strategy for the developmental literature review. Adapted from Page et al. [95].

venue and citation counts—two fairly reliable indicators of influence [7]. Top-priority papers include those published in first and second quartile journals for their respective discipline, as per the Scimago Journal Rank scheme, along with those published in conferences sponsored by the ACM and IEEE, given that these are the venues in which leading contributions are likely to be found [7, 126]. We make adjustments based on citation counts—as reported in the paper’s respective database—to identify central contributions that were published in lesser-known venues. This approach strategically delimits the number of papers included in the review while mitigating the bias towards highly cited publications or those published in prominent venues. At this stage, a total of 67 papers are chosen.

Finally, we perform full-text reviews of each of the 67 papers, excluding those that are irrelevant according to the initial inclusion criteria, extended abstracts, shorter than four pages, or elaborated further by the same authors in a subsequent study. Ultimately, a sample of 42 papers are included for analysis. Figure 1 presents these reasons for exclusion in a PRISMA diagram [95].

### 3.3 Analysis

Our data analysis employs both deductive and inductive elements. The deductive elements are the categories delineated in Table 3; these were defined prior to data extraction. Meanwhile, the inductive elements were defined during analysis according to the data; these are represented in Sections 5.1 through 5.3. We also extract bibliometric information, such as title, author(s), publication venue, abstract, and year. This scheme is coded into NVivo 12 to facilitate structured data extraction and analysis.

Throughout our analysis, we continuously review extracted segments: conceptually relevant extracts are grouped together and assigned an inductive code; these codes are added, combined, separated, or removed as further studies are analysed; and patterns among inductive codes are

Table 3. Data Extraction Form (Adapted from Bandara et al. [7])

Category	Description	RQ	Relevant Section
Title	Title of the paper	–	–
Author(s)	Author(s) of the paper	–	–
Publication venue	Journal or conference in which paper is published	–	Appendix B
Year	Year in which paper is published	Description	Section 4
Technology	Type of robot in which variable autonomy is implemented	Description	Section 4
Domain	Application domain in which the paper is situated	Description	Section 4
Definition	Definition of variable autonomy (or adjacent terms)	RQ1	Section 5.1
Motivation	Motivation or purpose for researching variable autonomy	RQ1	Section 5.1
Design	Paper's study design	RQ2	Section 5.2
Site	Empirical site in which study takes place	RQ2	Section 5.2
Evaluation	Paper's method to evaluate approach	RQ2	Section 5.2
Architecture	Technical description of variable autonomy implementation	RQ3	Section 5.3

identified to determine higher-level relationships to inform the development of our conceptual framework presented in Section 6 [7, 83].

It is worth offering further clarification on the Architecture category. Initially, we gather sub-codes from previous reviews [15, 82, 87]. We follow a flexible approach where new dimensions are added while some dimensions found in previous reviews are excluded. As an example, past reviews do not differentiate between changes in autonomy determined before operation or at runtime; our distinction between goal-oriented and stimulus-driven approaches captures this nuance. We expand on the similarities and differences between the dimensions of variable autonomy proposed in this article with past work in Section 6.

In summary, we aim to reconcile the conceptual and operational ambiguity around implementations of variable autonomy to devise an approach relevant for responsible robotics. With this aim in mind, we employ a developmental review of recent work in the variable autonomy literature. We leverage three search strategies to ensure a breadth of coverage, combined with a prioritisation strategy that delimited the corpus to a manageable number of prominent and representative publications. Finally, we employ an analysis approach that draws on deductive and inductive elements; the results of this analysis are presented in the following sections.

### 3.4 Limitations

We now deal with four limitations of our study. First, search queries are an inherently restrictive sampling strategy: only papers which use equivalent language will be returned as a result. Therefore, those which employ dissimilar language yet are still relevant will be excluded. We attempt to mitigate this risk by developing an extensive search query as shown in Appendix A. The terms in the query are gathered inductively by the first author from early papers and past reviews on variable autonomy; the search query was then reviewed by the second and third authors and revised accordingly. Additionally, we use multiple sampling strategies, such as forward and backward searches, to further offset this limitation.

Second, the process of data selection and analysis includes numerous decisions that may impact the internal validity of results. Therefore, we iteratively develop a data selection and extraction protocol. The data selection protocol is encoded in Microsoft Excel and the data extraction protocol is encoded in Nvivo 12 to support consistency.

Third, our search strategy draws from four sources of data: Scopus, WoS, IEEE Xplore, and ACM Digital Library. Although each of these databases index high impact conferences and journals, some relevant papers may be omitted. Nonetheless, the number of data sources in our review exceeds the minimum of 2 suggested by Shea et al. [113].

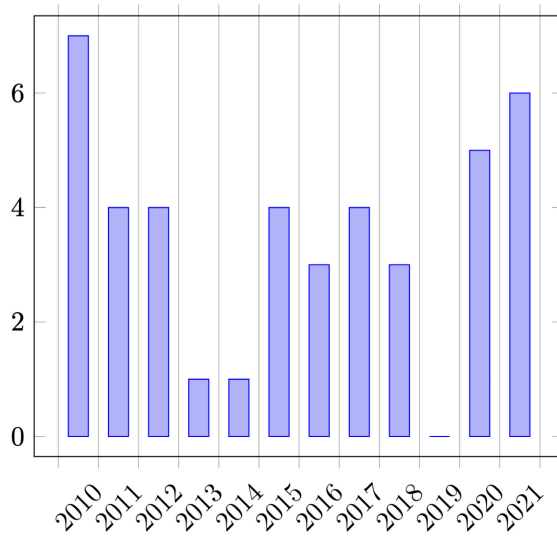


Fig. 2. Number of papers published by year (2010–2021).

Finally, we build our research agenda from what is currently possible from the perspective of technical research on variable autonomy. As such, research that does not focus on the design and implementation of variable autonomy is excluded from our search strategy. An implication of this choice is that studies which adopt a qualitative orientation to HRI and social robotics may not be included. In addition, although there is a productive community of scholarship that takes a qualitative approach to the study of human interactions with robots [e.g., 81, 142], as far as we are aware, such studies have not yet been extended to variable autonomy implementations.

#### 4 DATA DESCRIPTION

Our review includes 42 papers published in journals and conferences spanning from 2010 to 2021 and a diversity of application domains and robot technologies. The list of publication venues covered in this review is included in Appendix B. In this section, we present a brief description of our dataset. The intention of these statistics is not to infer properties of variable autonomy research in general, but to depict the breadth of publications included within our review.

As shown in Figure 2, the number of publications is fairly constant across the 12-year period between 2010 and 2021. Our dataset is evenly distributed, with half of the papers (21 of 42) published between 2010 and 2015, and the remaining half published between 2016 and 2021.

Figure 3 shows the application domains addressed in the reviewed papers. The most common are search and rescue (13 of 42) [2, 22, 25–27, 38, 39, 45, 73, 79, 102, 124, 131] and military (9 of 42) [30, 34, 39, 62, 88, 105, 119, 143, 144] contexts: the former refers to the use of robotics to identify and rescue missing persons in, for example, disaster scenarios; the latter includes the use of robotics for military operations such as surveillance, reconnaissance, and defence. Eight of 42 papers do not state a specific domain and are categorised as generic.

Six different types of robot technology are addressed in the reviewed papers, as shown in Figure 4. Sixteen studies include mobile robots [8, 25–28, 38, 70, 85, 101, 107, 111, 116, 124, 128, 131, 137], many of which are prototypes. Next, self-driving cars [8, 30, 34, 52, 100, 105, 119, 135, 144, 145]—autonomous ground vehicles that could transport passengers and cargo—and

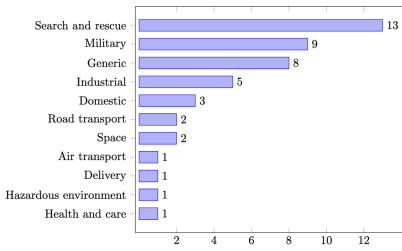


Fig. 3. Application domains addressed in the reviewed papers. The sum exceeds the number of papers reviewed due to papers discussing multiple domains.

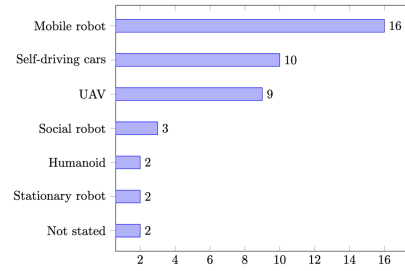


Fig. 4. Types of robot technology addressed in the reviewed papers. The sum exceeds the number of papers reviewed due to papers discussing multiple robot technologies.

**Unmanned Aerial Vehicles (UAVs)** [2, 22, 24, 31, 62, 73, 88, 101, 143] are featured in 10 and 9 papers, respectively.

## 5 RESULTS

In this section, we present our results framed as responses to each of our three research questions. First, we review common definitions of variable autonomy in the literature and distil their central features. We contextualise these definitions with the motivations for conducting variable autonomy research and present a comprehensive definition. Second, we describe the process of variable autonomy research, focusing on research designs, research sites, and evaluative measures. Third, we present a taxonomic representation of variable autonomy implementations across four dimensions, stated informally as questions: *who* initiates changes in autonomy, *what* aspects of autonomy are adjusted, *when* are changes determined, and *why* do changes occur? We provide formal characterisations of each dimension in turn.

### 5.1 RQ1: How Is Variable Autonomy Defined in the Literature?

As alluded to in Section 2.2, the literature on variable autonomy lacks consistent terminology. Different terms are given equivalent definitions, similar terms are alternatively defined, and some terms are given no definition at all. Further, there is no central definition to which authors commonly refer. Therefore, we propose a comprehensive definition that, when combined with the dimensions of variable autonomy discussed in Section 5.3, offers precision when describing robots with variable autonomy.

Of the 42 papers, the authors of 30 explicitly define their conceptualisation of variable autonomy, and across these 30 papers, six different terms appear. These terms, listed from highest to lowest number of appearances, include adjustable autonomy, adaptive autonomy, variable autonomy, sliding autonomy, adaptable autonomy, and dynamic autonomy. For some authors, the choice between these terms signals different approaches to variable autonomy. Adjustable and adaptable autonomy may represent systems in which changes in a robot’s autonomy are initiated by a human operator, whereas adaptive autonomy describes systems in which changes are triggered by the robot agent [51, 62, 131]. Valero-Gomez et al. [131] offer a representative distinction: “*adjustable autonomy*, in which the operator has initiative over the autonomy level; *adaptive autonomy*, in which the autonomy level is adjusted depending on the task and context” [emphasis in original] (p. 703). From this definition, we see that adjustments in autonomy are associated with particular conditions of the context of use and can be initiated by either a human or artificial agent.

Most of the papers which identify their approach as adaptive autonomy align with this distinction [2, 30, 45]; meanwhile, those that employ adjustable autonomy use the term much more

loosely. Specifically, these authors refer to adaptive and adjustable autonomy, along with other terms such as sliding autonomy, inconsistently or interchangeably [8, 22, 70, 73, 101]. For example, Basich et al. [8] define adjustable autonomy as “the ability of an autonomous system to alter its level of autonomy during plan execution, often by dynamically imposing or relaxing constraints on the extent of actions it can perform autonomously in a human-agent team” (p. 124). Similarly, Lewis et al. [70] refer to adjustable autonomy as “having the robots alter their level of autonomy in a situationally-dependent manner” (p. 1657). Next, an example of interchangeable use of terms, Roehr and Shi [101] state that “sliding autonomy also known as adaptive/adjustable autonomy and mixed initiative control is one area . . . [motivated by] increasing the efficiency of mixed [human-robot] teams by adjusting the autonomy level of individual robots” (p. 508). Here, sliding, adaptive, and adjustable are treated as equivalent terms, and the focus of the definition shifts to human-robot collaboration. These definitions complicate the adaptable/adjustable and adaptive autonomy distinction, and point out the dynamic nature of autonomy in variable autonomy systems.

The third most common term is variable autonomy, favoured by Chiou et al. [25–27, 100] and Ramesh et al. [100]. Chiou et al. [25] indicate that a “variable autonomy system is one in which control can be traded between a human operator and a robot by *switching between different Levels of Autonomy*” (p. 2). In comparison, this definition makes no claim as to *who* effects change; the emphasis is instead on *what* is changed.

Despite the inconsistent terminology, researchers’ motivations for pursuing variable autonomy are fairly similar. Researchers position variable autonomy as a strategy for groups comprised of both humans and robots to interact with one another, thereby balancing the strengths and limitations of autonomy with those of human operators. In particular, autonomous robot behaviour is seen to reduce operator workload, stress, and fatigue, and compensate for losses in an operator’s situation awareness: the ability to sense and perceive the robot’s operating environment [26, 30, 38, 100]. Human operators, however, are valued for their ability to respond to and navigate complex and uncertain environments [27, 70, 73, 87, 91, 116, 131]. Researchers, implicitly or explicitly, view this capability balancing as a means to improve the effectiveness, efficiency, and safety of the joint human-robot team [25, 31, 85, 88, 101, 107, 116]. Two papers offer an alternative framing, instead stating that the motivation for variable autonomy is to enable automation to adapt to the needs of human operators [51, 85].

From the preceding discussion and the results presented in Section 5.3, five fundamental concepts related to variable autonomy arise. The first two, LoAs and dynamism, are closely linked. In other words, the robot must possess multiple LoAs and the capacity to move between them during operation. Importantly, these changes can be initiated by either the human, robot, or both. Next, variable autonomy is an interaction strategy for groups comprised of both human and robot agents, each of whom possess distinct capabilities. As such, HRI considerations are central to the operationalisation of variable autonomy. Finally, changes in autonomy are deliberate: contextual cues trigger an adjustment from one LoA to another. Drawing together these concepts, we propose the following definition for variable autonomy in robotics.

*An interaction strategy between human and robot agents in which the robot’s level of autonomy varies during operation in response to changes in context.*

This definition makes explicit the five fundamental concepts of variable autonomy, whereas many of these are omitted from the reviewed definitions. Additionally, it includes both systems in which changes in autonomy are initiated by either the human, robot, or a combination of both; the intention is that this merging will remove unnecessary separation between related research efforts.

## 5.2 RQ2: How Is Research into Variable Autonomy Conducted?

In this section, we discuss three features of variable autonomy research: the research design employed, the research site, and measures used for evaluation. Reporting on the research design and site provides insight into the state of variable autonomy research, and, relatedly, the robustness of results. Depending on how results are generated and in what context they arise inferences can be made on their validity. Additionally, the measures researchers choose for evaluation and comparison reveal the qualities valued in variable autonomy implementations.

*5.2.1 Research Design.* Ordered from most to least common, variable autonomy researchers report on a range of research designs, as shown in Table 4: experimental, simulation, field tests, conceptual, and surveys. All research designs besides those categorised as conceptual or survey were task-oriented: a human-robot team, whether real or simulated, had to complete some pre-defined task.

Experimental designs refer to studies in which human participants act as a robot operator and perform a series of tasks under varying experimental conditions. Many experimental studies involve participants operating a robot across multiple LoAs while performing a secondary task, such as responding to questions [73] or mentally rotating three-dimensional objects [25]. Secondary tasks enable researchers to test operators' situation awareness [26] and induce cognitive load [25–27, 34, 85]. The participants in these studies constitute a relatively homogeneous population: 11 of 28 experiments rely on undergraduate and graduate students from the authors' respective universities [24, 25, 30, 34, 52, 73, 92, 131, 143, 144], and 5 of 27 employ members of the research team [23, 61, 88, 102, 128]. For 11 of 27 experimental papers, the participant sampling strategy is unclear [26, 27, 31, 38, 62, 70, 91, 111, 116, 124, 137], and 1 paper recruits participants from the lead author's research institution [105].

There are three variations of experimental design: within-subjects, between-subjects, and single-subject; it is unclear which approach is followed for six papers. The differences between these three refers to how many experimental conditions, or independent variables, each participant experiences. A within-subjects design has each participant experience each condition, whereas between-subjects exposes each participant to only one condition; for both within- and between-subjects, either one or multiple conditions can be tested. Most within-subjects experiments are single factor, meaning they only test one independent variable across each participant; these studies compare different implementations, such as teleoperation and variable autonomy [24, 38, 70], variable autonomy with other static LoAs [27, 92, 116, 124, 144], or systems in which changes in autonomy are triggered by the system or the human operator [25, 62, 143]. Meanwhile, the remaining within- and between-subjects studies test multiple independent variables, such as implementation (e.g., static vs. variable autonomy), operator and robot workload, and task difficulty [26, 30, 34, 73]. Three studies test unique conditions, such as differences in interfaces [105], alerts for changing autonomy [52], and number of robots [131]. Lastly, the single-subject designs imply that the study includes only one participant, a design used in preliminary work [23] or as a supplement to field tests [102].

Simulations, however, rely on numerical experiments within a virtual environment. For seven of nine papers that employ a simulation design, it serves as preliminary validation for a proposed variable autonomy architecture [8, 22, 28, 45, 101, 119, 135]. In contrast, Miller et al. [85] compare the predictive performance of different information streams for triggering shifts in LoAs, including signals from human control, autonomy, and the environment.

For the remaining papers, four report on field tests, such as in robotics competitions [79, 91, 102] or navigation through difficult terrain [107]; three papers introduce conceptual frameworks for performance measures to trigger adjustments in autonomy [2, 100, 145]; one paper presents the

Table 4. Research Designs Employed by Authors in the Reviewed Studies

Research Design	Occurrences	References
Simulation	9	[8, 22, 28, 45, 79, 85, 101, 119, 135]
Conceptual	3	[2, 100, 145]
Experimental		
<i>Between-subjects</i>	3	[26, 52, 131]
<i>Not stated</i>	6	[31, 61, 88, 111, 124, 137]
<i>Single-subject</i>	3	[23, 102, 128]
<i>Within-subjects</i>	16	[24, 25, 27, 30, 34, 38, 62, 70, 73, 91, 92, 105, 116, 124, 143, 144]
Field test	4	[79, 91, 102, 107]
Survey	1	[51]
Unclear	1	[39]

The sum exceeds the number of papers reviewed due to papers reporting on results from multiple research designs.

results of a survey that explores how older adults would respond to changes in a social robot's autonomy if triggered automatically or by the user; and for one paper it is unclear whether the results are from a simulation or experiment [39].

In summary, the majority of studies in this review employ an experimental design. Across these studies, the participants come from a limited subset of possible populations. Additionally, the experimental design varies from study to study, making it difficult to compare results.

**5.2.2 Research Site.** All four field tests and half of experiments (14/28) occur in physical research sites: arenas designed specifically for robotics trials [24, 79, 88, 91, 102], realistic outdoor [107] and indoor [38] settings, or, common for experiments, contrived environments such as indoor obstacles courses [23, 25, 26, 28, 61, 70, 85, 92, 116, 124, 128, 137]. The remaining half of experiments (14/28) take place in virtual environments, such as simulated programmes [24, 27, 30, 31, 34, 52, 62, 73, 105, 111, 124, 131, 143, 144].

These results show that most variable autonomy studies take place in artificial settings, whether in contrived physical environments or simulations. Variable autonomy implementations, therefore, are not evaluated in contexts that reflect the dynamism and complexities of the real world, a central factor motivating variable autonomy research.

**5.2.3 Evaluation Measures.** Researchers who conduct experimental studies and field tests employ an array of constructs and associated measures to evaluate variable autonomy implementations. Within the reviewed studies, constructs fall into two categories: *capability* constructs, which focus on the performance of either the operator or the robot in completing a pre-defined task, and *collaboration* constructs which characterise the quality of collaboration between the human and robot. Tables 5 and 6 detail the capability and collaboration constructs, respectively. Across each construct, measures are either objective or subjective, a common distinction in the HRI literature: the latter refers to measures that draw from the experiences and perceptions of the participant, commonly recorded through Likert-style surveys provided after the experiment; the former refers to data that is "independent" of the participant, recorded manually by the researcher or through devices such as sensors and timers.

Capability constructs include *effectiveness*, *efficiency*, *safety*, *situation awareness*, *adaptability*, *border-line functioning*, and *workload*. Objective measures of effectiveness and efficiency such as whether the primary task of operating the robot was successfully completed, the number of errors, and task completion time are ubiquitous. Many of these are idiosyncratic to each study, such as the number of targets accurately identified in a surveillance mission [34] or total area explored

Table 5. Capability Constructs and Associated Objective and Subjective Measures Focus on the Performance of Either the Operator or the Robot in Completing a Pre-Defined Task

Construct	Objective Measure(s)	Subjective Measure(s)
Adaptability	Acceptance rate of proposed autonomy changes	–
Border-line functioning	Time taken to resolve perturbation	–
Effectiveness	Number of errors Primary task success rate	–
Efficiency	Task completion time	–
Safety	Primary task success rate	–
Situation awareness	Secondary task success rate Amount of information exchange	Situation Awareness Rating Technique
Workload	Amount of information exchanged Operator energy expenditure Time spent in LoA	Mental demand Physical demand Operator performance Temporal demand Effort Task difficulty Frustration

in a search and rescue simulation [131]. For studies in which errors are associated with vehicle collisions, researchers interpret primary task success rate as a measure of safety [25–27, 38, 52]. Relatedly, Zieba et al. [144] employ two unique constructs of adaptability and border-line functioning, which refer to the ability of the system to manage issues and “border-line use conditions in a given operational mode” (p. 381), respectively.

Whereas measures of effectiveness and efficiency are exclusively objective, workload and situation awareness include measures both drawn from the operator’s experience and behavioural data. A common instrument for measuring subjective mental workload of task execution is the **NASA Task Load Index (NASA TLX)** method [25–27, 30, 34, 73, 85, 116, 124]. The NASA TLX is a well-established survey, composed of six dimensions: mental demand, physical demand, temporal demand (i.e., how rushed a participant felt), effort, performance, and frustration level. After completing a trial, participants rate their response to each question on a scale from low (1) to high (20). Although not all papers directly use the NASA TLX survey, some include closely related questions covering task difficulty [62] and perceived stress [38]. These are combined with objective measures of workload, including operator energy expenditure, calculated in terms of mechanical work [92]; amount of information exchanged between operator and robot [39, 79]; and time spent in each LoA [107]. Similarly, a combination of objective and subjective measures represent situation awareness. For example, Kidwell et al. [62] interpret participant performance on secondary tasks as an indication of situation awareness, whereas Côté et al. [31] infer situation awareness from the amount of environmental information displayed on a GUI throughout the duration of the experiment.

Constructs that refer to human-robot collaboration include *interaction effectiveness*, *interaction efficiency*, *automation reliance*, *trust*, *confidence*, and *acceptance*. The number of LoA switches [25,

Table 6. Collaboration Constructs and Associated Objective and Subjective Measures  
Characterise the Quality of Collaboration between the Human and Robot

Construct	Objective Measure(s)	Subjective Measure(s)
Acceptance	–	Preference Intention to use Perceived usefulness
Interaction effectiveness	Number of LoA switches Time spent in LoA	Comfort with automation
Interaction efficiency [8]	Number of HRIs Operator reaction time Time spent in LoA	–
Confidence	–	Robot performance Operator performance
Reliance	Number of LoA switches Time spent in LoA	–
Trust	Number of LoA switches Time spent in LoA	Robot performance

62, 101, 105] and time spent in each LoA [25, 52, 62, 91, 105, 131], along with the number of HRIs [31] and operator reaction time [143, 144], are collectively interpreted as reflecting the effectiveness and efficiency of interactions, how reliant participants are on automation, and the trust participants have in the robot. Owan et al. [92] evaluate participants' level of comfort engaging with the robot as a subjective measure of collaboration effectiveness. Similarly, three studies include questions to gauge participants' trust in automation [34, 62, 92]. Finally, measures of acceptance were mainly informal survey questions, asking participants to state their preferences between control modes [62, 73, 92, 116, 143], intention to use, and perceived usefulness [51].

The use and interpretation of measures varies significantly across the studies. For example, the number of LoA switches and time spent in each LoA is interpreted as an indicator of operator reliance on autonomy [105], trust [52], interaction efficiency [91, 101], and interaction effectiveness [131]. Even more, many studies do not explicitly state which constructs their measures are associated with. Additionally, the use of established subjective measures beyond the NASA TLX survey is limited. For other constructs such as situation awareness, trust, and acceptance, researchers rely on informal measures developed for the study at hand. Two exceptions are the experiments by de Visser and Parasuraman [34] and Owan et al. [92]: the former draws from the Situation Awareness Rating Technique by Taylor [125] and trust and self-confidence measure of Lee and Moray [69], and the latter adapts a questionnaire for human-robot collaboration fluency from Hoffman [53]. Finally, there are instances of joint use of objective and subjective measures to converge on a given construct. For example, Schaefer et al. [105] infer trust in automation through both the number of LoA switches and responses to trust questionnaires.

### 5.3 RQ3: How Is Variable Autonomy Implemented?

Implementations of variable autonomy differ across four dimensions, stated informally as questions: *who* initiates changes in autonomy, *what* aspects of autonomy are adjusted, *when* are changes determined, and *why* do changes occur? Each dimension includes several attributes. In this sec-

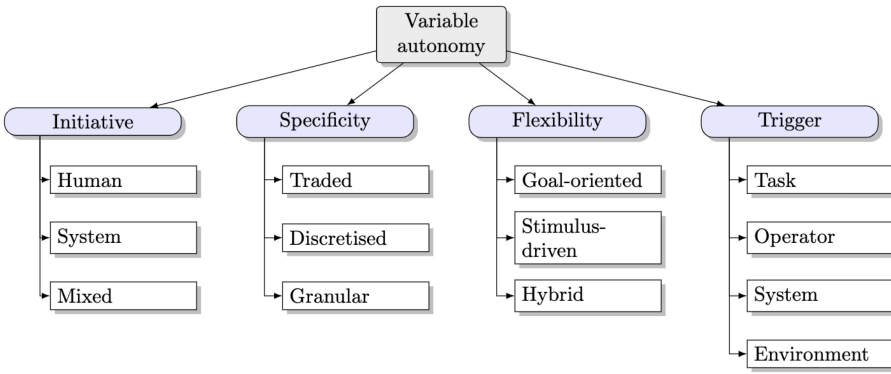


Fig. 5. Overview of the four dimensions of variable autonomy: initiative, specificity, flexibility, trigger.

tion, we detail the four dimensions in turn and describe the variety of considerations designers manage when constructing variable autonomy systems. An overview of the four dimensions and associated attributes is provided in Figure 5.

**5.3.1 Initiative.** A long-standing concern in variable autonomy is who initiates changes in autonomy. Whether it is the human, robot, or a combination of both represents our first dimension. We distinguish between these three types—human initiative, system initiative, and mixed initiative—and reflect on the implications of each.

*Human initiative* refers to implementations in which the human operator has sole capacity to change the robot’s autonomy. In one study, Lin and Goodrich [73] design an interface that enables an operator to manage the behaviour of a simulated UAV by setting the amount of time allocated to autonomy. In this instance, the human operator interprets information provided by the GUI to make a judgment on the appropriate level of autonomy during the task. Whereas the information provided by Lin and Goodrich’s interface is continuous, Bush et al. [22] present an architecture in which the robot issues a request for an autonomy switch based on the predicted likelihood of goal completion. Importantly, the robot could not initiate the change in autonomy itself and the operator could reject requests for assistance. Therefore, humans retain full control of changes in autonomy for human initiative variable autonomy systems, and may receive information that guides their decision on when to initiate a change either through continuously available information on an interface or discrete alerts sent by the robot. Besides serving as a medium for information on when to intervene, interface design also influences a human operator’s propensity to initiate autonomy changes. Schaefer et al. [105] find that operators are more likely to adjust a robot’s autonomy when the interface is familiar; drawing on past work in automation reliance, the authors suggest that the familiarity of interfaces mediates human trust in and reliance on robots.

Rather than relying on a human operator to adjust a robot’s autonomy, *system initiative* implementations enable the robot’s autonomy to change automatically. Specifically, a “control switcher” [25]—an artificial agent, such as a learning algorithm [38, 61, 70, 92, 119], fuzzy controller [111], Markov decision process [135, 137], or finite state machine [79]—adjusts the robot’s autonomy. For example, Doroodgar et al. [38] develop a hierarchical reinforcement learning algorithm that allocates control for performing a task to either a human operator or robot according to whichever agent is predicted to do so more efficiently. These systems obviate the need for human intervention in autonomy switches, yet still require human involvement. A transition from autonomous behaviour to teleoperation demands the availability and awareness of a human operator who is

willing and able to assume a greater degree of control following a period of passivity. Research on self-driving cars discusses the risk of “vigilance decrement” on behalf of operators when they remain in a passive state for an extended period of time [64].

Finally, *mixed initiative* implementations integrate the previous two types: both human operator and control switcher are able to initiate autonomy changes [25, 88, 100, 144]. The operator and robot must collaborate to determine the appropriate level of robot autonomy, with the most capable either seizing or being granted control [25]. As characterised by Chiou et al. [26], this implies that both the robot and human must have an understanding of the other’s state, knowledge, and capabilities. Recent experimental work by Chiou et al. [25] finds that mixed initiative systems improve performance and operator workload during navigation tasks as compared to human initiative systems, at least in a simulated environment.

**5.3.2 Specificity.** When developing a variable autonomy system, designers must specify what aspects of autonomy are subject to variation. Approaches found in the literature adjust autonomy between two or more discrete operation modes, or at a granular level of control for autonomous behaviour.

*Traded control* approaches shift between two extremes: manual and autonomous control [24, 26, 119]. A concern for this approach is that operators lose situation awareness during periods of inattention, and struggle to regain control after the robot’s autonomous behaviour decreases [52]. Cosenzo et al. [30] attempt to mitigate this risk by continuously reengaging the operator. Similarly, *discretised control* implementations include pre-defined LoAs with intermediate degrees of autonomy [28, 31, 85, 143]. As no studies in the review compare traded and discretised control implementations, there lacks evidence on the tradeoffs associated with employing either approach.

*Granular control* implementations do not conceptualise operation modes in terms of discrete LoAs. Instead, they adjust autonomy by constraining or expanding the functions a robot and human are allowed to do, required to do, and able to do [73, 128, 144, 145]. The continuous scale approach requires designers to exercise greater specificity in defining what autonomous behaviours will be adjusted—for example, Lin and Goodrich [73] set constraints on where a UAV could operate under autonomous behaviour and for how long.

**5.3.3 Flexibility.** A variable autonomy system is one in which the robot’s autonomy changes during operation. Some variable autonomy implementations provide greater flexibility in the number and timing of these adjustments than others. Our third dimension differentiates between systems in which changes in autonomy are defined *a priori* or occur dynamically.

In *goal-oriented* variable autonomy systems, when and what autonomy changes occur are defined before operation. In a study by Small et al. [116], the authors introduce a goal-oriented variable autonomy system, termed *Assigned Responsibility*, in which various segments of a task are assigned an LoA before operation, and the robot monitors the progress of task completion to automatically change LoAs as it moves from one segment to the next. This approach imposes rigidity on the system, but, as Small et al. [116] suggest, reduces the operator’s cognitive load and enables designers to explicitly state when automation will be used to align with legal and ethical considerations.

*Stimulus-driven* autonomy adjustments imply that all decisions related to changes in autonomy take place at runtime [2, 22, 24, 28, 31]. The human operator or control switcher dynamically adjusts autonomy during task execution, without following a prescribed set of changes. These approaches enable greater flexibility and the ability to respond to unpredictable circumstances, but introduce a degree of uncertainty in robot behaviour.

Of course, the choice is not binary. Some implementations, such as that proposed by Romay et al. [102] and Mostafa et al. [88], adopt a *hybrid* approach, in which designers define a relative LoA

for various task segments during the design stage while the operator retains the ability to make adjustments on-the-fly.

**5.3.4 Trigger.** According to our definition for variable autonomy, autonomy adjustments occur because of some change in context. Influenced by previous taxonomies for triggers in adaptive systems [43, 97], we organise triggers for variable autonomy systems into four categories: task, operator, system, and environment.

*Task triggers* address aspects of the task which the human-robot team performs, relying either on a measurement of the task's state or the properties ascribed to individual tasks by designers. Variable autonomy systems calculate task state indicators such as completion status [116, 124] and predicted likelihood of failure [22, 101]. In the goal-oriented approach by Small et al. [116], the system monitors task progress to automatically change LoA as the robot moves from one task to the next. Task completion is represented as an observable state of the world to which a current state is continuously compared against. These triggers require the system to sense its surrounding environment and relate environmental conditions to the ongoing task. Another grouping of task triggers address properties of the task itself: some studies distinguish between types of tasks, labelling some as sensitive and therefore requiring human, rather than autonomous, control [102, 137]; others switch between control modes as the relative difficulty of a task changes [30, 34, 39, 88, 143]. For example, de Visser and Parasuraman [34] develop a system initiative architecture that moves from manual to autonomous control as task load, defined in terms of the number of vehicles under an operator's supervision, increases. Similarly, Mostafa et al. [88] develop a system that varies its autonomy according to a task's complexity, calculated by the number of individual actions required to complete it.

*Operator triggers* reflect the states and decisions of the human operator. Several studies attempt to infer internal properties such as operator workload through physiological sensors [143] and competence level through the amount and quality of human input [70, 85, 111]. Zhao et al. [143] employ eye trackers and sensor-enabled wristbands to measure cognitive processing and stress levels, whereas Lewis et al. [70] develop a model of expert-novice differences to increase the degree of autonomy when lower-skilled operators engage with the system. Whereas such systems require the ability to sense aspects of the operator, others defer to an operator's own judgment. Some studies indicate that an operator's judgment on when to adjust a robot's autonomy may be influenced by individual characteristics such as personality, preferences, trust, and experience with robots [25–27, 62, 101].

*System triggers* refer to events and states internal to the robot. There are two varieties of system triggers: monitoring and error detection. The difference between the two is one of severity: monitoring approaches measure gradual changes in system performance, whereas error detection focuses on discrete failures in autonomy. By comparing current to expected performance, monitoring techniques initiate changes in autonomy whenever system performance falls below a given threshold [24, 25, 61, 119]. A recent paper by Ramesh et al. [100] proposes the concept of “robot vitals,” a composite measure of performance in multi-robot systems; the vitals include “rate of change of signal strength, sliding window average of difference between expected robot velocity and actual velocity, robot acceleration, rate of increase in area coverage, and localisation error” (p. 303). The authors argue that the relative simplicity of their measure supports explainability in a robot's decisions. Meanwhile, error detection triggers changes whenever the autonomy fails [52, 91, 107, 124, 144].

Finally, *environment triggers* capture the circumstances of the robot's external environment. For example, a robot may enter a manual control mode when entering a novel environment or encountering unforeseen events [79]. Likewise, changing environmental conditions such as weather and

obstacles may require an operator to take or relinquish control from the autonomy [92, 107, 135]. Robots must be able to sense their surrounding environment for these triggers to function.

## 6 DISCUSSION: DESIGN GUIDELINES FOR VARIABLE AUTONOMY THROUGH RESPONSIBLE ROBOTICS

We reviewed 42 recent papers on variable autonomy to investigate how variable autonomy is defined in the literature (RQ1), how research into variable autonomy is conducted (RQ2), and how variable autonomy is implemented (RQ3). Overall, our review makes four contributions. First, we provide a definition of variable autonomy synthesised from past definitions in the literature. As shown by our results in Section 5.1, the variable autonomy literature employs diverse and inconsistent terminology and definitions. We attempt to clarify the field's language by offering a synthesised definition that builds on past articulations and incorporates the four dimensions of variable autonomy.

Second, we detail the research designs, sites, and measures employed in the literature to support rigorous empirical research. We provide evidence for the concern that the results of variable autonomy research may lack ecological validity given that most studies have been conducted in artificial settings, such as simulations or contrived physical environments [87]. As such, these studies have not been evaluated in contexts that reflect the dynamism and complexities of the real world. Additionally, we highlight how variable autonomy research follows a restrictive definition of relevant stakeholders, focusing only on the role of the operator rather than any other implicated group such as bystanders or passengers. Further, we point to the field's limited modes of evaluation; most empirical studies rely on homegrown measures, rather than utilising validated instruments, and do not include qualitative evidence surrounding people's experiences with variable autonomy robots. These challenges are not restricted to variable autonomy but have been noted in the field of HRI more broadly [33, 63].

Third, we distil previous characterisations of variable autonomy to provide a heuristic for designers when defining requirements for variable autonomy robotics. In particular, we deepen the description of the triggers that initiate changes in autonomy and introduce the dimension of "flexibility" to distinguish between implementations that allow for changes in autonomy to be determined before operation or at runtime. Further, previous reviews include several dimensions which we argue are not specific to variable autonomy but are relevant to autonomy more broadly; these include human-agent interaction, autonomy representation, and autonomy measurement [87]. Therefore, our taxonomy offers a concise formulation of aspects that distinguish variable autonomy from other HRI strategies.

Finally, in this section, we draw inspiration from Jirotko et al. [58] and Amershi et al. [4] to present 11 design guidelines (DG1–DG11) that will help researchers approach variable autonomy through a lens of responsible robotics. These guidelines, depicted in Table 7, touch upon the product and process of innovation, as introduced in Section 2.3, and build upon the results from our review in Section 5.

*DG1: Select Ethical Robotics Principles.* There are several resources that outline ethical principles for robotics (see [140]). Select one as a basis for ethical reflection throughout the duration of the research and innovation process while remaining flexible so the principles can be adapted to fit the circumstances of project stakeholders.

*DG2: Determine the Objectives of the Robotic System.* As shown in the discussion of researcher motivations and evaluation measures, the values underpinning variable autonomy research are predominantly performance-based. The concern is how to enable a human and robot to interact with one another to achieve some objective. Yet, a responsible robotics approach to variable au-

Table 7. Eleven Design Guidelines for Variable Autonomy Research Based on Responsible Robotics

Design Guidelines	
Process	1 Select ethical robotics principles.
	2 Determine the objectives of the robotic system.
	3 Identify relevant stakeholders beyond users.
	4 Conduct ethical risk assessment.
	5 Sample representative participants from stakeholder populations.
	6 Create study design with stakeholder input.
	7 Develop holistic evaluation plan.
Product	8 Match initiative to context.
	9 Support specific control modes.
	10 Enable flexible autonomy changes.
	11 Select appropriate triggers.

tonomy entails a wider range of goals, such as supporting stakeholder physical and psychological well-being and minimising environmental harm.

*DG3: Identify Relevant Stakeholders beyond Users.* Stakeholders are “those who are or will be significantly implicated by the technology” [46, p. 35]. A stakeholder can be one who directly interacts with a technology or one who does not interact with it but is still impacted by its use—a distinction between direct and indirect stakeholders, respectively. Within variable autonomy research, most participants assume the role of operators. This presents an abstraction of how robots would be used in practical contexts—for example, there are networks of different humans who exist in the robot’s operating environment. The IEEE 7001-2021 standard for “Transparency of Autonomous Systems”[117] includes several categories of direct and indirect stakeholders to consider, such as non-expert users, domain expert users, superusers, the general public, and bystanders. Take the scenario of an assistive robot within a care home: the intended user may be an older adult with support needs, but she does not live in isolation. She is likely supported by a network of family members, friends, care workers, and physicians. Each of these groups may have separate experiences of and responses to the use robots with variable autonomy.

*DG4: Conduct Ethical Risk Assessment.* The British Standard 8611 (BS8611) outlines a systematic approach to identify, analyse, and mitigate ethical hazards associated with the design and application of robots [55]. It includes a taxonomy of 20 ethical hazards that designers can draw from to reduce the effect of ethical harms—that is, harms that compromise psychological, societal, or environmental well-being.

*DG5: Sample Representative Participants from Stakeholder Populations.* As shown in this review, research on variable autonomy—and HRI more generally [63]—relies on non-representative groups, namely university students and members of the research team, to act as prospective robot operators. These groups may not actually display the same characteristics as future relevant stakeholders given differences in age and professional history. Therefore, the preferences and attitudes towards robots expressed by these study populations may not represent those of other populations.

*DG6: Create Research Design with Stakeholder Input.* Collaborate with stakeholders to determine where the study will be conducted, the tasks to be performed, how different types of stakeholders will be included, and whether the approach is acceptable. From this process, researchers should clearly specify the research design employed. For example, when following an experimental setup,

researchers should articulate whether it follows a between-, within-, or single-subject(s) design; the independent variable(s); the evaluative measures (along with what construct each is meant to operationalise); and the research site. Additionally, researchers should use this as an opportunity to extend beyond the traditional experimental paradigm, towards studies that focus on “how real people, in real-world environments, would interact face to face with a real robot” [33]. In other words, research should evaluate variable autonomy implementations in contexts that reflect the dynamism and complexities of the real world.

*DG7: Develop Holistic Evaluation Plan.* Evaluation should employ well-established quantitative and qualitative methods. For quantitative evaluation, rather than developing homegrown objective and subjective measures, leverage validated instruments for constructs such as workload (see [50]), situation awareness (see [40, 125]), trust (see [68, 69, 75]), psychological safety (see [67]), and usability (see [71]). There are numerous reviews that outline common measures used in HRI research (e.g., [29, 32, 78, 89, 121]). Although quantitative evaluation allows for comparison across individuals and the potential for generalisable knowledge, it misses out on the meanings and values people ascribe to phenomena within specific contexts. Qualitative methods such as interviews [10, 11] and ethnography [57] enable researchers to engage with such concepts.

*DG8: Match Initiative to Context.* Deciding who has the authority to initiate changes in autonomy has implications for the performance of the human-robot team, the experience of the human operator, as well as the experiences of other people who either directly or indirectly interact with the robot. The choice between human, system, and mixed initiative implementations entails trade-offs between factors such as human control, efficiency, and consistency, and therefore should be made in relation to the context in which the robot will be used.

*DG9: Support Specific Control Modes.* LoAs are a useful construct to help us understand variation in autonomous capabilities. However, in actual implementations, different autonomous capabilities are allocated to different functions and may change depending on the activity [16]. Therefore, greater specificity of choice in autonomous capabilities, such as in discretised and granular control, enables the robot and human to fine-tune autonomous capabilities to the current situation.

*DG10: Enable Flexible Autonomy Changes.* The flexibility of the variable autonomy implementation concerns the designer’s ability to specify *a priori* the types of behaviour which will be performed under certain control modes. Goal-oriented approaches enable designers to pre-define the allocation of autonomous capabilities. Defining exactly when the robot will operate with certain autonomous capabilities is useful in regulated or safety-critical contexts where the use of autonomy to perform certain tasks may be restricted. Yet, this regulation of autonomous behaviour increases its rigidity, and may preclude the operator or the robot from adapting dynamically in uncertain and unforeseen situations. Dynamic adjustments in autonomy, as stated before, imply greater variability and unpredictability in behaviour: an operator may be unprepared to regain control when it is handed back to her, or she may retake it when she is not suited to perform the task at hand. Hybrid approaches, therefore, provide a middle-ground route where certain behaviours can be assigned to autonomous capabilities beforehand while retaining the system and/or operator’s ability to make adjustments as changes in context arise.

*DG11: Select Appropriate Triggers.* Responding to changes in context requires the use of sensing capabilities. These triggers, such as those that infer an operator’s state or environmental conditions, may introduce privacy and security concerns depending on the type of data collected. Audio and video data of the operating environment may capture personal information if used in a sensitive context such as a person’s home. Data collected within search and rescue and

military applications may depict traumatising experiences or confidential national security details. Decisions on the types of triggers should be made on a case-by-case basis, as the operating environment determines the data that is likely to be collected. Different jurisdictions face different regulatory requirements for data collection and processing, and these should serve as a foundation for these decisions.

An important property of these design guidelines is that they are not speculative; several of these recommendations have already been successfully applied on numerous robotics projects. First, we begin with the process-oriented design guidelines (DG1–DG7). In a project on accidents involving autonomous vehicles, Ten Holter et al. [127] describe how they based their approach on the AREA framework (DG1) and drew on the expertise of stakeholders such as insurers, scholars, engineers, pedestrians, and cycling groups to inform their research plan (DG3, DG6). Meanwhile, McGinn et al. [80] build on BS8611 to conduct an ethical assessment of a real-world disinfectant robot used in a hospital in Ireland (DG4). Moving towards the design guidelines focused on product (DG8–DG11), our recommendations have been drawn from the technical literature: Small et al. [116] point to the utility of system initiative architectures in predictable environments (DG8), the work of Lin and Goodrich [73] introduces an innovative strategy to enable specific modes of autonomy adjustment (DG9), Romay et al. [102] and Mostafa et al. [88] enable flexible autonomy changes (DG10), and Ramesh et al. [100] propose a unique set of performance monitoring measures that support explainability in a robot's autonomy adjustment decisions (DG11).

## 7 FUTURE WORK

This article's objective is to establish a research agenda for variable autonomy based on responsible robotics. The relationship between these two areas is in its early stages and has yet to be investigated through primary research. Therefore, we propose the following research agenda.

*Responsibility.* As discussed, there are two notions of responsibility: forward- and backward-looking [133]. Forward-looking responsibility depends on the anticipation of consequences. Therefore, we ask what concerns and challenges stakeholders anticipate regarding the use of variable autonomy robotics, particularly across different design configurations. This inquiry will seek to provide an empirical basis for our initial explorations of impacts discussed in this section and to conceptualise how variable autonomy design features can mitigate the adverse consequences of robotics in varied contexts. Next, backward-looking responsibility requires the ability to assess a past sequence of events. We are exploring the concept of an **Ethical Black Box (EBB)**, a device similar to a flight data recorder that continuously records sensor inputs, actuator outputs, and relevant internal status data to facilitate accident investigations involving robots (see the work of Winfield et al. [141] and Winfield and Jirotko [139] for further discussion on the concept of an EBB). As such, we are interested in how variable autonomy can be incorporated into EBB recordings and how relevant information can be interpreted during accident investigations.

*Product, Process, Purpose, and People of Innovation.* Responsible robotics implies a concern for a broad range of stakeholders, moving beyond the perspective of the individual user [74, 122, 132]. Yet, our results show that variable autonomy research has thus far only been concerned with the perspective of the operator; therefore, we will explore how diverse stakeholders experience variable autonomy. Specifically, we are interested in how variable autonomy affects stakeholders' trust and confidence in a robot, as well as their sense of autonomy and identity, using both qualitative data and quantitative measures. Additionally, we will ask whether these experiences and perceptions vary across contexts, such as application domains and technology.

*Societal Values.* When taking a values-led approach to design, there is a balance to be struck between relying on *a priori* values and those that are situated and local. As such, we pick up two directions in our investigations of values and variable autonomy. First, how can we use variable autonomy to enable robots to adapt to the values of different cultural contexts? Second, the values currently driving variable autonomy research, as inferred from the researchers' motivations and evaluation measures, are predominantly performance-based. Therefore, we will explore holistic approaches to the evaluation of variable autonomy that span beyond concerns of efficiency and effectiveness.

## 8 CONCLUSION

In this article, we conducted a developmental review on variable autonomy in robotics to establish the foundation for a research agenda in responsible robotics. To conclude, we summarise this work's key findings:

- We define variable autonomy in robotics as “an interaction strategy between human and robot agents in which the robot's level of autonomy varies during operation in response to changes in context.”
- Based on the research designs and sites of empirical studies, most variable autonomy implementations have not been evaluated in contexts that reflect the dynamism and complexities of the real world, a central factor motivating variable autonomy research.
- From the motivations discussed in the papers and the evaluation measures employed in empirical studies, variable autonomy research is driven by performance-based values, such as efficiency and effectiveness in relation to some task. Additionally, the perspectives of human participants are reduced to questions of mental workload, preference, and technology acceptance through instruments such as the NASA TLX survey. There is a need to conduct more holistic evaluation of variable autonomy implementations that combines a diversity of quantitative measures and qualitative explorations of participants' experiences.
- Participants included in empirical studies are limited in terms of representativeness and diversity. First, most study participants are either university students or members of the research team, populations that may not represent relevant stakeholder groups. Second, the only stakeholder role investigated across the reviewed studies is that of the operator, whereas a robot in a practical context is likely to interact with networks of different humans.
- Variable autonomy implementations vary along four dimensions: initiative (the agent that initiates changes in the robot's autonomy), specificity (the aspects of autonomy that are changed), flexibility (the point at which autonomy changes are determined), and trigger (the contextual features that stimulate changes in autonomy).

## APPENDICES

### A DATABASE SEARCH QUERIES

The search queries used for the four databases—Scopus, WoS, ACM Digital Library, and IEEE Xplore—are shown in Table 8.

Table 8. Search Queries for Scopus, WoS, ACM Digital Library, and IEEE Xplore Databases

Database	Query	Results
Scopus	(variable OR adjustable OR adaptive OR dynamic OR flexible OR sliding) PRE/0 autonomy) AND robot* OR self?driving OR ((autonomous OR unmanned OR uninhabited) PRE/1 vehicle))	99
WoS	TS=((variable OR adjustable OR adaptive OR dynamic OR flexible OR sliding ) NEAR/0 autonomy) AND TS=(robot* OR self?driving OR ((autonomous OR unmanned OR uninhabited) NEAR/1 vehicle))	84
ACM Digital Library	[[All: “variable autonomy”] OR [All: “adjustable autonomy”] OR [All: “adaptive autonomy”] OR [All: “dynamic autonomy”] OR [All: “flexible autonomy”] OR [All: “sliding autonomy”]] AND [[All: robot*] OR [All: self?driving] OR [[[All: autonomous] OR [All: unmanned] OR [All: uninhabited]] AND [All: vehicle]]]	49
IEEE Xplore	((("All Metadata":variable OR "All Metadata":adjustable OR "All Metadata":adaptive OR "All Metadata":dynamic OR "All Metadata":flexible OR "All Metadata":sliding) ONEAR/1 "All Metadata":autonomy) AND ("All Metadata":robot* OR "All Metadata":self?driving OR ("All Metadata":autonomous OR "All Metadata":unmanned OR "All Metadata":uninhabited) ONEAR/2 "All Metadata":vehicle)))	62

All searches were conducted in January 2022, and results were restricted to the years 2010 – 2021.

B PUBLICATION VENUES

The publication venues for variable autonomy papers included in this review are shown in Table 9.

Table 9. List of Publication Venues Included in the Review

Publication Venue	No. of Papers
IEEE International Conference on Intelligent Robots and Systems	5
Proceedings of the Human Factors and Ergonomics Society	3
IEEE Robotics and Automation Letters	2
IEEE/WIC/ACM International Conference on Intelligent Agent Technology	1
Chinese Journal of Aeronautics	1
International Joint Conference on Autonomous Agents and Multiagent Systems	1
IEEE Aerospace Conference Proceedings	1
Transportation Research Part C	1
IEEE Intelligent Systems	1
International Conference on Collaboration Technologies and Systems, CTS	1
IEEE International Conference on Enabling Technologies: Infrastructures for Collaborative Enterprises	1
International Journal of Intelligent Robotics and Applications	1
ACM/IEEE International Conference on Human-Robot Interaction	1
Sensors	1
IEEE International Conference on Robotics and Automation	1
Cognition, Technology, and Work	1
IEEE International Conference on Systems, Man, and Cybernetics	1
Information Sciences	1
IEEE International Symposium on Computational Intelligence and Informatics, CINTI	1
International Conference on Self-Adaptive and Self-Organizing Systems, SASO	1
IEEE International Workshop on Robot and Human Interactive Communication	1
International Journal of Intelligent Computing and Cybernetics	1
International Symposium on Artificial Intelligence, Robotics and Automation in Space	1
Human Factors: The Journal of the Human Factors and Ergonomics Society	1
Journal of Cognitive Engineering and Decision Making	1
Journal of Field Robotics	1
Journal of Intelligent and Robotic Systems: Theory and Applications	1
Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems	1
IEEE SOUTHEASTCON	1
Robotica	1
IEEE-RAS International Conference on Humanoid Robots	1
Springer Tracts in Advanced Robotics	1
IEEE/ACM Workshop on AI Engineering–Software Engineering for AI (WAIN)	1
ACM Transactions on Human-Robot Interaction	1
IEEE/ASME International Conference on Advanced Intelligent Mechatronics	1

REFERENCES

[1] David A. Abbink, Tom Carlson, Mark Mulder, Joost C. F. de Winter, Farzad Aminravan, Tricia L. Gibo, and Erwin R. Boer. 2018. A topology of shared control systems—Finding common ground in diversity. *IEEE Transactions on Human-Machine Systems* 48, 5 (2018), 509–525. <https://doi.org/10.1109/THMS.2018.2791570>

[2] Sophia Abraham, Zachariah Carmichael, Sreya Banerjee, Rosaura VidalMata, Ankit Agrawal, Md. Nafee Al Islam, Walter Scheirer, and Jane Cleland-Huang. 2021. Adaptive autonomy in human-on-the-loop vision-based robotics systems. In *Proceedings of the 2021 IEEE/ACM 1st Workshop on AI Engineering–Software Engineering for AI (WAIN ’21)*. IEEE, Los Alamitos, CA, 113–120. <https://doi.org/10.1109/WAIN52551.2021.00025>

[3] Julie A. Adams, Pramila Rani, and Nilanjan Sarkar. 2004. Mixed initiative interaction and robotic systems. In *Proceedings of the AAAI Workshop on Supervisory Control of Learning and Adaptive Systems*. 6–13.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19)*. ACM, New York, NY, 1–13. <https://doi.org/10.1145/3290605.3300233>

[5] Lisanne Bainbridge. 1983. Ironies of automation. *Automatica* 19, 6 (1983), 775–779.

[6] Michael Baker and Holly A. Yanco. 2004. Autonomy mode suggestions for improving human-robot interaction. In *Proceedings of the 2004 IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 3. IEEE, Los Alamitos, CA, 2948–2953.

- [7] Wasana Bandara, Suraya Miskon, and Erwin Fieft. 2011. A systematic, tool-supported method for conducting literature reviews in information systems. In *Proceedings of the 19th European Conference on Information Systems (ECIS '11)*. 15.
- [8] Connor Basich, Justin Svegliato, Kyle Hollins Wray, Stefan Witwicki, Joydeep Biswas, and Shlomo Zilberstein. 2020. Learning to optimize autonomy in competence-aware systems. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '20)*. 123–131.
- [9] Jenay M. Beer, Arthur D. Fisk, and Wendy A. Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction* 3, 2 (2014), 74. <https://doi.org/10.5898/JHRI.3.2.Beer>
- [10] Cindy L. Bethel, Jessie E. Cossitt, Zachary Henkel, and Kenna Baugus. 2020. Qualitative interview techniques for human-robot interactions. In *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer Series on Bio- and Neurosystems, Vol. 12. Springer, 145–174.
- [11] Cindy L. Bethel, Zachary Henkel, and Kenna Baugus. 2020. Conducting studies in human-robot interaction. In *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer Series on Bio- and Neurosystems, Vol. 12. Springer, 91–124.
- [12] Vincent Blok and Pieter Lemmens. 2015. The emerging concept of responsible innovation. Three reasons why it is questionable and calls for a radical transformation of the concept of innovation. In *Responsible Innovation 2*. Springer, 19–35.
- [13] Marianne Boenink and Olya Kudina. 2020. Values in responsible research and innovation: From entities to practices. *Journal of Responsible Innovation* 7, 3 (2020), 450–470.
- [14] Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1125–1134. <https://doi.org/10.1145/2207676.2208560>
- [15] Jeffrey M. Bradshaw, Paul J. Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. 2004. Dimensions of adjustable autonomy and mixed-initiative interaction. In *Agents and Computational Autonomy*. Lecture Notes in Computer Science, Vol. 2969. Springer, 17–39. [https://doi.org/10.1007/978-3-540-25928-2\\_3](https://doi.org/10.1007/978-3-540-25928-2_3)
- [16] Jeffrey M. Bradshaw, Robert R. Hoffman, David D. Woods, and Matthew Johnson. 2013. The seven deadly myths of “autonomous systems.” *IEEE Intelligent Systems* 28, 3 (2013), 54–61. <https://doi.org/10.1109/MIS.2013.70>
- [17] Jeffrey M. Bradshaw, Maarten Sierhuis, Alessandro Acquisti, Paul Feltovich, Robert Hoffman, Renia Jeffers, Debbie Prescott, Niranjani Suri, Andrzej Uszok, and Ron Van Hoof. 2003. Adjustable autonomy and human-agent teamwork in practice: An interim report on space applications. In *Agent Autonomy*, Henry Hexmoor, Cristiano Castelfranchi, and Rino Falcone (Eds.). Multiagent Systems, Artificial Societies, and Simulated Organizations, Vol. 7. Springer US, 243–280. [https://doi.org/10.1007/978-1-4419-9198-0\\_11](https://doi.org/10.1007/978-1-4419-9198-0_11)
- [18] Jeffrey M. Bradshaw, Maarten Sierhuis, Yuri Gawdiak, Renia Jeffers, Niranjani Suri, and Mark Greaves. 2001. Adjustable autonomy and teamwork for the personal satellite assistant. In *Proceedings of the IJCAI-01 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents*. 1–6.
- [19] Martim Brandão, Masoumeh Mansouri, and Martin Magnusson. 2022. Editorial: Responsible robotics. *Frontiers in Robotics and AI* 9 (2022), 937612. <https://doi.org/10.3389/frobt.2022.937612>
- [20] J. Brookshire, S. Singh, and R. Simmons. 2004. Preliminary results in sliding autonomy for assembly by coordinated teams. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '04)*, Vol. 1. IEEE, Los Alamitos, CA, 706–711. <https://doi.org/10.1109/IROS.2004.1389435>
- [21] David J. Bruemmer, Donald D. Dudenhofer, and Julie L. Marble. 2002. Dynamic-autonomy for urban search and rescue. In *Proceedings of the AAAI Mobile Robot Competition*. 33–37.
- [22] L. A. M. Bush, A. J. Wang, and B. C. Williams. 2012. Risk-based sensing in support of adjustable autonomy. In *Proceedings of the 2012 IEEE Aerospace Conference*. IEEE, Los Alamitos, CA, 1–18. <https://doi.org/10.1109/AERO.2012.6187312>
- [23] Samuel Bustamante, Gabriel Quere, Katharina Hagmann, Xuwei Wu, Peter Schmaus, Jorn Vogel, Freck Stulp, and Daniel Leidner. 2021. Toward seamless transitions between shared control and supervised autonomy in robotic assistance. *IEEE Robotics and Automation Letters* 6, 2 (2021), 3833–3840. <https://doi.org/10.1109/LRA.2021.3064449>
- [24] Jonathan Cacace, Alberto Finzi, and Vincenzo Lippiello. 2014. A mixed-initiative control system for an aerial service vehicle supported by force feedback. In *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Los Alamitos, CA, 1230–1235. <https://doi.org/10.1109/IROS.2014.6942714>
- [25] Manolis Chiou, Nick Hawes, and Rustam Stolkin. 2021. Mixed-initiative variable autonomy for remotely operated mobile robots. *ACM Transactions on Human-Robot Interaction* 10, 4 (2021), 1–34. <https://doi.org/10.1145/3472206>
- [26] Manolis Chiou, Nick Hawes, Rustam Stolkin, Kimron L. Shapiro, Jess R. Kerlin, and Andrew Clouter. 2015. Towards the principled study of variable autonomy in mobile robots. In *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, Los Alamitos, CA, 1053–1059. <https://doi.org/10.1109/SMC.2015.190>
- [27] Manolis Chiou, Rustam Stolkin, Goda Bieksaite, Nick Hawes, Kimron L. Shapiro, and Timothy S. Harrison. 2016. Experimental analysis of a variable autonomy framework for controlling a remotely operating mobile robot. In

- Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '16)*. IEEE, Los Alamitos, CA, 3581–3588. <https://doi.org/10.1109/IROS.2016.7759527>
- [28] David C. Conner and Justin Willis. 2017. Flexible navigation: Finite state machine-based integrated navigation and control for ROS enabled robots. In *Proceedings of SoutheastCon 2017*. IEEE, Los Alamitos, CA, 1–8. <https://doi.org/10.1109/SECON.2017.7925266>
- [29] Enrique Coronado, Takuya Kiyokawa, Gustavo A. Garcia Ricardez, Ixchel G. Ramirez-Alpizar, Gentiane Venture, and Natsuki Yamanobe. 2022. Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an Industry 5.0. *Journal of Manufacturing Systems* 63 (2022), 392–410. <https://doi.org/10.1016/j.jmmsy.2022.04.007>
- [30] Keryl Cosenzo, Jessie Chen, Lauren Reinerman-Jones, Michael Barnes, and Denise Nicholson. 2010. Adaptive automation effects on operator performance during a reconnaissance mission with an unmanned ground vehicle. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 54 (2010), 2135–2139.
- [31] Nicolas Côté, Arnaud Canu, Maroua Bouzid, and Abdel-lillah Mouaddib. 2012. Humans-robots sliding collaboration control in complex environments with adjustable autonomy. In *Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE, Los Alamitos, CA, 146–153. <https://doi.org/10.1109/WI-IAT.2012.215>
- [32] Praveen Damacharla, Ahmad Y. Javaid, Jennie J. Gallimore, and Vijay K. Devabhaktuni. 2018. Common metrics to benchmark human-machine teams (HMT): A review. *IEEE Access* 6 (2018), 38637–38655. <https://doi.org/10.1109/ACCESS.2018.2853560>
- [33] Kerstin Dautenhahn. 2018. Some brief thoughts on the past and future of human-robot interaction. *Journal of Human-Robot Interaction* 7, 1 (May 2018), Article 4, 3 pages. <https://doi.org/10.1145/3209769>
- [34] Ewart de Visser and Raja Parasuraman. 2011. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making* 5, 2 (2011), 209–231. <https://doi.org/10.1177/1555343411410160>
- [35] Ewart J. de Visser, Melanie LeGoullon, Amos Freedy, Elan Freedy, Gershon Weltman, and Raja Parasuraman. 2008. Designing an adaptive automation system for human supervision of unmanned vehicles: A bridge from theory to practice. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52 (2008), 221–225.
- [36] M. Desai and H. A. Yanco. 2005. Blending human and robot inputs for sliding scale autonomy. In *Proceedings of the 2005 International Workshop on Robot and Human Interactive Communication (ROMAN '05)*. IEEE, Los Alamitos, CA, 537–542. <https://doi.org/10.1109/ROMAN.2005.1513835>
- [37] Gregory A. Dorais, R. Peter Bonasso, David Kortenkamp, Barney Pell, and Debra Schreckenghost. 1999. *Adjustable Autonomy for Human-Centered Autonomous Systems*. Technical Report. Scientific Research Publishing.
- [38] Barzin Doroodgar, Maurizio Ficocelli, Babak Mobedi, and Goldie Nejat. 2010. The search for survivors: Cooperative human-robot interaction in search and rescue environments using semi-autonomous robots. In *Proceedings of the 2010 IEEE International Conference on Robotics and Automation*. IEEE, Los Alamitos, CA, 2858–2863. <https://doi.org/10.1109/ROBOT.2010.5509530>
- [39] Domagoj Drenjanac, Slobodanka Dana Kathrin Tomic, and Eva Kuhn. 2015. A semantic framework for modeling adaptive autonomy in task allocation in robotic fleets. In *Proceedings of the 2015 IEEE 24th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*. IEEE, Los Alamitos, CA, 15–20. <https://doi.org/10.1109/WETICE.2015.29>
- [40] Mica R. Endsley. 2017. Direct measurement of situation awareness: Validity and use of SAGAT. In *Situational Awareness*. Routledge, Milton Park, UK, 129–156.
- [41] Mica R. Endsley. 2017. From here to autonomy: Lessons learned from human-automation research. *Human Factors* 59, 1 (2017), 5–27. <https://doi.org/10.1177/0018720816681350>
- [42] EPSRC. 2022. Anticipate, Reflect, Engage, Act (AREA). Retrieved December 15, 2023 from <https://www.ukri.org/about-us/epsrc/our-policies-and-standards/framework-for-responsible-innovation/>
- [43] Karen M. Feigh, Michael C. Dorneich, and Caroline C. Hayes. 2012. Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 54, 6 (2012), 1008–1024. <https://doi.org/10.1177/0018720812443983>
- [44] David Feil-Seifer and Maja J. Mataric. 2011. Socially assistive robotics: Ethical issues related to technology. *IEEE Robotics & Automation Magazine* 18, 1 (2011), 24–31. <https://doi.org/10.1109/MRA.2010.940150>
- [45] Mirgita Frasheri, Baran Curuklu, Mikael Esktrom, and Alessandro Vittorio Papadopoulos. 2018. Adaptive autonomy in a search and rescue scenario. In *Proceedings of the 2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO '18)*. IEEE, Los Alamitos, CA, 150–155. <https://doi.org/10.1109/SASO.2018.00026>
- [46] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, Cambridge, MA.

- [47] M. A. Goodrich, Dan R. Olsen, Jacob W. Crandall, and Thomas J. Palmer. 2001. Experiments in adjustable autonomy. In *Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics: e-Systems and e-Man for Cybernetics in Cyberspace*, Vol. 3. IEEE, Los Alamitos, CA, 1624–1629. <https://doi.org/10.1109/ICSMC.2001.973517>
- [48] David J. Gunkel. 2020. Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology* 22, 4 (2020), 307–320.
- [49] Benjamin Hardin and Michael A. Goodrich. 2009. On using mixed-initiative control: A perspective for managing large-scale robotic teams. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction (HRI '09)*. ACM, New York, NY, 165. <https://doi.org/10.1145/1514095.1514126>
- [50] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [51] Marcel Heerink. 2011. How elderly users of a socially interactive robot experience adaptiveness, adaptability and user control. In *Proceedings of the 2011 IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI '11)*. IEEE, Los Alamitos, CA, 79–84. <https://doi.org/10.1109/CINTI.2011.6108476>
- [52] Michelle Hester, Kevin Lee, and Brian P. Dyre. 2017. “Driver take over”: A preliminary exploration of driver trust and performance in autonomous vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (2017), 1969–1973. <https://doi.org/10.1177/1541931213601971>
- [53] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.
- [54] Hui-Min Huang, Kerry Pavek, James Albus, and Elena Messina. 2005. Autonomy levels for unmanned systems (AL-FUS) framework: An update. In *Unmanned Ground Vehicle Technology VII*, Vol. 5804. SPIE, 439–448.
- [55] British Standards Institute. 2016. *BS8611:2016 Robots and Robotic Devices, Guide to the Ethical Design and Application of Robots and Robotic Systems*. British Standards Institute.
- [56] SAE International. 2021. *SAE J3016 202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. SAE International.
- [57] An Jacobs, Shirley A. Elprama, and Charlotte I. C. Jewell. 2020. Evaluating human-robot interaction with ethnography. *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer Series on Bio- and Neurosystems, Vol. 12. Springer, 269–286.
- [58] Marina Jirotk, Barbara Grimpe, Bernd Stahl, Grace Eden, and Mark Hartswood. 2017. Responsible research and innovation in the digital age. *Communications of the ACM* 60, 5 (2017), 62–68.
- [59] David B. Kaber and Mica R. Endsley. 2004. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science* 5, 2 (2004), 113–153. <https://doi.org/10.1080/1463922021000054335>
- [60] David B. Kaber, Jennifer M. Riley, Kheng-Wooi Tan, and Mica R. Endsley. 2001. On the design of adaptive automation for complex systems. *International Journal of Cognitive Ergonomics* 5, 1 (2001), 37–57. [https://doi.org/10.1207/S15327566IJCCE0501\\_3](https://doi.org/10.1207/S15327566IJCCE0501_3)
- [61] Uri Kartoun, Helman Stern, and Yael Edan. 2010. A human-robot collaborative reinforcement learning algorithm. *Journal of Intelligent & Robotic Systems* 60, 2 (2010), 217–239. <https://doi.org/10.1007/s10846-010-9422-y>
- [62] Brian Kidwell, Gloria L. Calhoun, Heath A. Ruff, and Raja Parasuraman. 2012. Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1 (2012), 428–432. <https://doi.org/10.1177/1071181312561096>
- [63] Sara Kiesler and Michael A. Goodrich. 2018. The science of human-robot interaction. *Journal of Human-Robot Interaction* 7, 1 (May 2018), Article 9, 3 pages. <https://doi.org/10.1145/3209701>
- [64] Moritz Körber, Andrea Cingel, Markus Zimmermann, and Klaus Bengler. 2015. Vigilance decrement and passive fatigue caused by monotony in automated driving. *Procedia Manufacturing* 3 (2015), 2403–2409.
- [65] David Kortenkamp, R. Peter Bonasso, Dan Ryan, and Debbie Schreckenghost. 1997. Traded control with autonomous robots as mixed initiative interaction. In *Proceedings of the AAAI Symposium on Mixed Initiative Interaction*. 89–94.
- [66] David Kortenkamp, D. Keirn-Schreckenghost, and R. P. Bonasso. 2000. Adjustable control autonomy for manned space flight. In *Proceedings of the 2000 IEEE Aerospace Conference*, Vol. 7. 629–640. <https://doi.org/10.1109/AERO.2000.879330>
- [67] Przemysław A. Lasota, Terrence Fong, and Julie A. Shah. 2017. A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics* 5, 4 (2017), 261–349.
- [68] Theresa Law and Matthias Scheutz. 2021. *Trust: Recent Concepts and Evaluations in Human-Robot Interaction*. Academic Press, 27–57.
- [69] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270. <https://doi.org/10.1080/00140139208967392>
- [70] Bennie Lewis, Bulent Tastan, and Gita Sukthankar. 2013. An adjustable autonomy paradigm for adapting to expert-novice differences. In *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Los Alamitos, CA, 1656–1662. <https://doi.org/10.1109/IROS.2013.6696571>

- [71] James R. Lewis. 2018. The system usability scale: Past, present, and future. *International Journal of Human-Computer Interaction* 34, 7 (2018), 577–590.
- [72] Shuhong Lla, A. V. Wynsberghe, and Sabine Roeser. 2020. The complexity of autonomy: A consideration of the impacts of care robots on the autonomy of elderly care receivers. In *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020*. Frontiers in Artificial Intelligence and Applications, Vol. 335: Culturally Sustainable Social Robotics. IOS Press, 316–325.
- [73] Lanny Lin and Michael A. Goodrich. 2015. Sliding autonomy for UAV path-planning: Adding new dimensions to autonomy management. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '15)*. 10.
- [74] Rob Lubberink, Vincent Blok, Johan Van Ophem, and Onno Omta. 2017. Lessons for responsible innovation in the business context: A systematic literature review of responsible, social and sustainable innovation practices. *Sustainability* 9, 5 (2017), 721.
- [75] Bertram F. Malle and Daniel Ullman. 2021. *A Multidimensional Conception and Measure of Human-Robot Trust*. Academic Press, 3–25.
- [76] J. L. Marble, D. J. Bruemmer, and D. A. Few. 2003. Lessons learned from usability tests with a collaborative cognitive workspace for human-robot teams. In *Proceedings of the 2003 IEEE International Conference on Systems, Man, and Cybernetics (SMC '03)*, Vol. 1. IEEE, Los Alamitos, CA, 448–453. <https://doi.org/10.1109/ICSMC.2003.1243856>
- [77] Julie L. Marble, David J. Bruemmer, Douglas A. Few, and Donald D. Dudenhofer. 2004. Evaluation of supervisory vs. peer-peer interaction for human-robot teams. In *Proceedings of the 37th Annual Hawaii International Conference on Systems Sciences*. 1–22.
- [78] Jeremy A. Marvel, Shelly Bagchi, Megan Zimmerman, and Brian Antonishek. 2020. Towards effective interface designs for collaborative HRI in manufacturing: Metrics and measures. *Journal of Human-Robot Interaction* 9, 4 (May 2020), Article 25, 55 pages. <https://doi.org/10.1145/3385009>
- [79] Stephen McGill, Seung-Joon Yi, and Daniel D. Lee. 2015. Team THOR's adaptive autonomy for disaster response humanoids. In *Proceedings of the 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids '15)*. IEEE, Los Alamitos, CA, 453–460. <https://doi.org/10.1109/HUMANOIDS.2015.7363589>
- [80] Conor McGinn, Robert Scott, Niamh Donnelly, Michael F. Cullinan, Alan Winfield, and Pat Treusch. 2023. Ethical assessment of a hospital disinfection robot. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 12008–12014.
- [81] Antonia Meissner, Angelika Trübswetter, Antonia S. Conti-Kufner, and Jonas Schmidler. 2020. Friend or foe? Understanding assembly workers' acceptance of human-robot collaboration. *Journal of Human-Robot Interaction* 10, 1 (2020), Article 3, 30 pages. <https://doi.org/10.1145/3399433>
- [82] Leila Methnani, Andrea Aler Tubella, Virginia Dignum, and Andreas Theodorou. 2021. Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence* 4 (2021), 737072. <https://doi.org/10.3389/frai.2021.737072>
- [83] Matthew B. Miles, A. Micheal Huberman, and Saladaña. 2019. *Qualitative Data Analysis: A Methods Sourcebook* (4th ed.). SAGE Publications.
- [84] Christopher A. Miller and Raja Parasuraman. 2007. Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49, 1 (2007), 57–75. <https://doi.org/10.1518/001872007779598037>
- [85] Christopher X. Miller, Temesgen Gebrekristos, Michael Young, Enid Montague, and Brenna Argall. 2021. An analysis of human-robot information streams to inform dynamic autonomy allocation. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '21)*. IEEE, Los Alamitos, CA, 1872–1878. <https://doi.org/10.1109/IROS51168.2021.9636637>
- [86] John Mingers and Loet Leydesdorff. 2015. A review of theory and practice in scientometrics. *European Journal of Operational Research* 246, 1 (2015), 1–19. <https://doi.org/10.1016/j.ejor.2015.04.002>
- [87] Salama A. Mostafa, Mohd Sharifuddin Ahmad, and Aida Mustapha. 2019. Adjustable autonomy: A systematic literature review. *Artificial Intelligence Review* 51, 2 (2019), 149–186. <https://doi.org/10.1007/s10462-017-9560-8>
- [88] Salama A. Mostafa, Mohd Sharifuddin Ahmad, Aida Mustapha, and Mazin Abed Mohammed. 2017. Formulating layered adjustable autonomy for unmanned aerial vehicles. *International Journal of Intelligent Computing and Cybernetics* 10, 4 (2017), 430–450. <https://doi.org/10.1108/IJICC-02-2017-0013>
- [89] Robin R. Murphy and Debra Schreckenghost. 2013. Survey of metrics for human-robot interaction. In *Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13)*. 197–198. <https://doi.org/10.1109/HRI.2013.6483569>
- [90] David Musliner and Barney Pell. 1999. *Agents with Adjustable Autonomy: Papers from the 1999 AAAI Symposium, March 22–24, Stanford, California*. AAAI Press.

- [91] Sebastian Muszynski, Jorg Stuckler, and Sven Behnke. 2012. Adjustable autonomy for mobile teleoperation of personal service robots. In *Proceedings of the 2012 21st IEEE International Symposium on Robot and Human Interactive Communication (ROMAN '12)*. IEEE, Los Alamitos, CA, 933–940. <https://doi.org/10.1109/ROMAN.2012.6343870>
- [92] Parker Owan, Joseph Garbini, and Santosh Devasia. 2020. Faster confined space manufacturing teleoperation through dynamic autonomy with task dynamics imitation learning. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2357–2364. <https://doi.org/10.1109/LRA.2020.2970653>
- [93] Richard Owen. 2014. The UK Engineering and Physical Sciences Research Council’s commitment to a framework for responsible innovation. *Journal of Responsible Innovation* 1, 1 (2014), 113–117.
- [94] Thomas O’Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human-autonomy teaming: A review and analysis of the empirical literature. *Human Factors* 64, 5 (2022), 904–938. <https://doi.org/10.1177/0018720820960865>
- [95] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery* 88 (2021), 105906.
- [96] Raja Parasuraman. 2000. Designing automation for human use: Empirical studies and quantitative models. *Ergonomics* 43, 7 (2000), 931–951.
- [97] Raja Parasuraman, Toufik Bahri, John E. Deaton, Jeffrey G. Morrison, and Michael Barnes. 1992. *Theory and Design of Adaptive Automation in Aviation Systems*. Technical Report. Cognitive Science Lab, Catholic University of America, Washington, DC.
- [98] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 30, 3 (2000), 286–297.
- [99] Jari Pirhonen, Helinä Melkas, Arto Laitinen, and Satu Pekkarinen. 2020. Could robots strengthen the sense of autonomy of older people residing in assisted living facilities?—A future-oriented study. *Ethics and Information Technology* 22, 2 (2020), 151–162.
- [100] Aniketh Ramesh, Manolis Chiou, and Rustam Stolkin. 2021. Robot vitals and robot health: An intuitive approach to quantifying and communicating predicted robot performance degradation in human-robot teams. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, 303–307. <https://doi.org/10.1145/3434074.3447181>
- [101] Thomas M. Roehr and Yuping Shi. 2010. Using a self-confidence measure for a system-initiated switch between autonomy modes. In *Proceedings of the 10th International Symposium on Artificial Intelligence, Robotics, and Automation in Space*. 507–514.
- [102] Alberto Romay, Stefan Kohlbrecher, Alexander Stumpf, Oskar von Stryk, Spyros Maniatopoulos, Hadas Kress-Gazit, Philipp Schillinger, and David C. Conner. 2017. Collaborative autonomy between high-level behaviors and human operators for remote manipulation tasks using different humanoid robots. *Journal of Field Robotics* 34, 2 (2017), 333–358. <https://doi.org/10.1002/rob.21671>
- [103] P. Scerri, David V. Pynadath, and Milind Tambe. 2002. Towards adjustable autonomy for the real world. *Journal of Artificial Intelligence Research* 17 (2002), 171–228. <https://doi.org/10.1613/jair.1037>
- [104] Paul Scerri. 2001. *Designing Agents for Systems with Adjustable Autonomy*. Ph.D. Dissertation. Linköping University.
- [105] Kristin E. Schaefer, Ralph W. Brewer, Joe Putney, Edward Mottern, Jeffrey Barghout, and Edward R. Straub. 2016. Relinquishing manual control: Collaboration requires the capability to understand robot intent. In *Proceedings of the 2016 International Conference on Collaboration Technologies and Systems (CTS '16)*. IEEE, Los Alamitos, CA, 359–366. <https://doi.org/10.1109/CTS.2016.0071>
- [106] René von Schomberg (Ed.). 2011. *Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields*. Publications Office of the European Union, Luxembourg.
- [107] Debra Schreckenghost, Tod Milam, and Terrence Fong. 2010. Measuring performance in real time during remote human-robot operations with adjustable autonomy. *IEEE Intelligent Systems* 25, 5 (2010), 36–45. <https://doi.org/10.1109/MIS.2010.126>
- [108] Deborah Scott. 2021. Diversifying the deliberative turn: Toward an agonistic RRI. *Science, Technology, & Human Values* 48, 2 (2021), 1–24.
- [109] B. Sellner, F. W. Heger, L. M. Hiatt, R. Simmons, and S. Singh. 2006. Coordinated multiagent teams and sliding autonomy for large-scale assembly. *Proceedings of the IEEE* 94, 7 (2006), 1425–1444. <https://doi.org/10.1109/JPROC.2006.876966>
- [110] Mario Selvaggio, Marco Cognetti, Stefanos Nikolaidis, Serena Ivaldi, and Bruno Siciliano. 2021. Autonomy in physical human-robot interaction: A brief survey. *IEEE Robotics and Automation Letters* 6, 4 (2021), 7989–7996. <https://doi.org/10.1109/LRA.2021.3100603>
- [111] Piatan Sfair Palar, Vinicius de Vargas Terres, and André Schneider de Oliveira. 2020. Human-robot interface for embedding sliding adjustable autonomy methods. *Sensors* 20, 20 (2020), 5960. <https://doi.org/10.3390/s20205960>

- [112] Amanda Sharkey and Noel Sharkey. 2012. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology* 14, 1 (2012), 27–40.
- [113] Beverley J. Shea, Jeremy M. Grimshaw, George A. Wells, Maarten Boers, Neil Andersson, Candyce Hamel, Ashley C. Porter, Peter Tugwell, David Moher, and Lex M. Bouter. 2007. Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology* 7, 1 (2007), 10–17.
- [114] Thomas B. Sheridan and William L. Verplank. 1978. *Human and Computer Control of Undersea Teleoperators*. Technical Report. Man-Machine Systems Lab, MIT, Cambridge, MA.
- [115] Ben Shneiderman. 2022. *Human-Centered AI*. Oxford University Press, Oxford, UK.
- [116] Nicolas Small, Kevin Lee, and Graham Mann. 2018. An assigned responsibility system for robotic teleoperation control. *International Journal of Intelligent Robotics and Applications* 2, 1 (2018), 81–97. <https://doi.org/10.1007/s41315-018-0043-0>
- [117] IEEE Computer Society. 2021. *IEEE Standard for Transparency of Autonomous Systems*. IEEE, Los Alamitos, CA.
- [118] IEEE Computer Society. 2021. *IEEE Standard Model Process for Addressing Ethical Concerns during System Design: IEEE Standard 7000-2021*. IEEE, Los Alamitos, CA.
- [119] Boris Sofman, J. Andrew Bagnell, and Anthony Stentz. 2010. Bandit-based online candidate selection for adjustable autonomy. In *Field and Service Robotics*, Andrew Howard, Karl Iagnemma, and Alonzo Kelly (Eds.). Vol. 62. Springer Tracts in Advanced Robotics, Vol. 62. Springer, 239–248. [https://doi.org/10.1007/978-3-642-13408-1\\_22](https://doi.org/10.1007/978-3-642-13408-1_22)
- [120] Bernd Carsten Stahl and Mark Coeckelbergh. 2016. Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems* 86 (2016), 152–161.
- [121] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI '06)*. ACM, New York, NY, 33–40. <https://doi.org/10.1145/1121241.1121249>
- [122] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (2013), 1568–1580. <https://doi.org/10.1016/j.respol.2013.05.008>
- [123] Araz Taeiagh and Hazel Si Min Lim. 2019. Governing autonomous vehicles: Emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews* 39, 1 (2019), 103–128.
- [124] Fang Tang, Mahmood Mohammed, and Jacob Longazo. 2016. Experiments of human-robot teaming under sliding autonomy. In *Proceedings of the 2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM '16)*. IEEE, Los Alamitos, CA, 113–118. <https://doi.org/10.1109/AIM.2016.7576752>
- [125] R. M. Taylor. 1989. Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations*.
- [126] Mathieu Templier and Guy Paré. 2015. A framework for guiding and evaluating literature reviews. *Communications of the Association for Information Systems* 37, 6 (2015), 112–137. <https://doi.org/10.17705/1CAIS.03706>
- [127] Carolyn Ten Holter, Lars Kunze, Jo-Ann Pattinson, and Marina Jirotko. 2022. Responsible innovation; responsible data. A case study in autonomous driving. *Journal of Responsible Technology* 11 (2022), 100038. <https://doi.org/10.1016/j.jrt.2022.100038>
- [128] C. Ton, Z. Kan, and S. S. Mehta. 2018. Obstacle avoidance control of a human-in-the-loop mobile robot system using harmonic potential fields. *Robotica* 36, 4 (2018), 463–483. <https://doi.org/10.1017/S0263574717000510>
- [129] Sherry Turkle. 2011. *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, New York, NY.
- [130] Marialena Vagia, Aksel A. Transeth, and Sigurd A. Fjerdingen. 2016. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics* 53 (2016), 190–202.
- [131] Alberto Valero-Gomez, Paloma de la Puente, and Miguel Hernando. 2011. Impact of two adjustable-autonomy models on the scalability of single-human/multiple-robot teams for exploration missions. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 6 (2011), 703–716. <https://doi.org/10.1177/0018720811420427>
- [132] Ibo van de Poel, Lotte Asveld, Steven Flipse, Pim Klaassen, Zenlin Kwee, Maria Maia, Elvio Mantovani, Christopher Nathan, Andrea Porcari, and Emad Yaghmaei. 2020. Learning to do responsible innovation in industry: Six lessons. *Journal of Responsible Innovation* 7, 3 (2020), 697–707.
- [133] Ibo Van de Poel and Martin Sand. 2021. Varieties of responsibility: Two problems of responsible innovation. *Synthese* 198, 19 (2021), 4769–4787.
- [134] Jeroen Van den Hoven. 2013. Value sensitive design and responsible innovation. In *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, Richard Owen, John Bessant, and Maggy Heintz (Eds.). Wiley, 75–83.
- [135] Franco van Wyk, Anahita Khojandi, and Neda Masoud. 2020. Optimal switching policy between driving entities in semi-autonomous vehicles. *Transportation Research Part C: Emerging Technologies* 114 (2020), 517–531. <https://doi.org/10.1016/j.trc.2020.02.011>

- [136] Aimee van Wynsberghe and Noel Sharkey. 2020. Special issue on responsible robotics: Introduction. *Ethics and Information Technology* 22 (2020), 281–282. <https://doi.org/10.1007/s10676-020-09562-y>
- [137] Lois Vanhee, Laurent Jeanpierre, and Abdel-Allah Mouaddib. 2021. Optimizing requests for support in context-restricted autonomy. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '21)*. IEEE, Los Alamitos, CA, 6434–6440. <https://doi.org/10.1109/IROS51168.2021.9636240>
- [138] Jijun Wang and Michael Lewis. 2007. Human control for cooperating robot teams. In *Proceeding of the ACM/IEEE International Conference on Human-Robot Interaction (HRI '07)*. ACM, New York, NY, 9. <https://doi.org/10.1145/1228716.1228719>
- [139] Alan F. T. Winfield and Marina Jirotko. 2017. The case for an ethical black box. In *Proceedings of the Annual Conference Towards Autonomous Robotic Systems*. 262–273.
- [140] Alan F. T. Winfield and Marina Jirotko. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 1–13.
- [141] Alan F. T. Winfield, Katie Winkle, Helena Webb, Ulrik Lyngs, Marina Jirotko, and Carl Macrae. 2021. Robot accident investigation: A case study in responsible robotics. In *Software Engineering for Robotics*. Springer, 165–187.
- [142] Katie Winkle, Praminda Caleb-Solly, Ailie Turton, and Paul Bremner. 2018. Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In *Proceedings of the 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. 289–297.
- [143] Zhe Zhao, Yifeng Niu, and Lincheng Shen. 2020. Adaptive level of autonomy for human-UAVs collaborative surveillance using situated fuzzy cognitive maps. *Chinese Journal of Aeronautics* 33, 11 (2020), 2835–2850. <https://doi.org/10.1016/j.cja.2020.03.031>
- [144] Stéphane Zieba, Philippe Polet, and Frédéric Vanderhaegen. 2011. Using adjustable autonomy and human-machine cooperation to make a human-machine system resilient—Application to a ground robotic system. *Information Sciences* 181, 3 (2011), 379–397. <https://doi.org/10.1016/j.ins.2010.09.035>
- [145] Stéphane Zieba, Philippe Polet, Frédéric Vanderhaegen, and Serge Debernard. 2010. Principles of adjustable autonomy: A framework for resilient human-machine cooperation. *Cognition, Technology & Work* 12, 3 (2010), 193–203. <https://doi.org/10.1007/s10111-009-0134-7>

Received 10 October 2022; revised 30 June 2023; accepted 9 November 2023