A responsibility-centered approach to defining levels of automation

IOP Conf. Series: Journal of Physics: Conf. Series **1357** (2019) 012027 doi:10.1088/1742-6596/1357/1/012027

A responsibility-centered approach to defining levels of automation

Bård Myhre[1,3], Are Hellandsvik[2] and Stig Petersen[2]

[1] SINTEF Digital, Oslo, Norway
[2] SINTEF Digital, Trondheim, Norway
[3] bard.myhre@sintef.no

Abstract. Several authors and organisations have over the last 40 years suggested various ways to describe different levels of automation. In this paper we give an introduction to some of the most relevant models and present some of the challenges faced when using them. Furthermore, we suggest a new definition for autonomy, based on responsibility instead of functionality. The paper analyses some relevant cases based on the suggested definition of autonomy, and points towards some issues that should be investigated in the maritime domain related to current definitions for autonomous ships.

1. Introduction

The concept of *autonomy* has received much interest and attention over the last decades and is by many regarded as a natural next step in the evolution of automation. Traditionally, the word *autonomous* has often been used as a synonym to *unmanned*, and with this interpretation autonomous systems have existed for at least 60 years through satellites, industrial robots and military weapon systems [1]. However, an exact and generally accepted definition of *autonomy* does not yet exist, suggesting some caution when investigating the topic.

Sheridan and Verplank [2] are often cited as some of the first addressing computer-made decisions in a systematic way, presenting ten levels of automation in their report on "Human and Computer Control of Undersea Teleoperators". This report also describes how the human and computer could cooperate on the various levels, adding that other variations are possible. Later, several other variants of *Levels of Automation (LOA)* have been suggested, typically with somewhat fewer levels. Examples are Endsley's four role allocations between expert system and pilot from 1987 [3][4], BASt's five categories of automated driving functions from 2013 [5], SAE's six levels of driving automation levels for on-road vehicles first published in 2014 and revised in 2016 and 2018 [6], and the four operational autonomy levels for autonomous merchant ships proposed by Norwegian Forum for Autonomous Ships (NFAS) in 2017 [7]. Of these, SAE's six levels seem to have gained the most solid foothold, as it is now being used by the U.S. Department of Transportation [8] and aims for ISO standardisation of its next revision [9].

A common denominator in all the above-mentioned models is the approach of categorizing current and future technologies according to their technical capabilities. However, lacking a common definition of autonomy, we argue that attempting to define automation levels in this manner risks being a theoretical discussion on how "smart" a system must be to be "autonomous". This risk is also mentioned by the US Naval Studies Board in their 2005 report on "Autonomous Vehicles in Support of Naval Operations" [1], and their pragmatic solution was to investigate unmanned systems in general instead of defining autonomy. This does not mean that pursuing an unambiguous definition of autonomy is of no value, but rather that we may benefit from a different approach. On this background

we suggest using the concept of *human responsibility* as a basis for the definition instead of *technical capability*. This shift in vantage point enables us to define what we think is a valid and relevant definition of *autonomy*, as well as describing distinct *automation levels* that may have implications on how we should understand, develop and use autonomous systems.

In order to establish an overview of the topic, section 2 gives a description of the most relevant Levels of Automation (LOA) models, their main characteristics and limitations. In section 3 we propose a new, responsibility-based definition of autonomy and how this definition can be used to define new levels of automation. Section 4 follows up with some examples on how the suggested definition of autonomy could be used for analyzing and understanding different cases of automation. In section 5 we indicate how a responsibility-centered definition of autonomy can be incorporated into existing frameworks for autonomous ships, while section 6 concludes the paper and suggests topics for further research.

2. Current models for Levels of Automation

As mentioned previously, there exist several variants of models for Levels of Automation (LOA). Here we present three of these in more detail; the original model by Sheridan and Verplank [2], the SAE model [6] and the NFAS model [7]. What they all have in common is that they describe how an automatic function (or computer) incrementally makes more decisions for a human operator, until it finally handles all situations by itself.

2.1. Sheridan and Verplank's levels of automation in man-computer decision-making

Sheridan and Verplank's original levels of automation are presented in table 1, as described in [2]. Note that a more clear-cut version presented in [10] is often used instead of the original model, but the two variants are in principle identical.

Table 1. Sheridan and Verplank's original levels of automation [2]

Level	Description of interaction
1	Human does the whole job up to the point of turning it over to the computer to implement.
2	Computer helps by determining the options.
3	Computer helps determine options and suggests one, which human need not follow.
4	Computer selects action and human may or may not do it.
5	Computer selects action and implements it if human approves.
6	Computer selects action, informs human in plenty of time to stop it.
7	Computer does the whole job and necessarily tells human what it did.
8	Computer does whole job and tells human what it did only if human explicitly asks.
9	Computer does whole job and tells human what it did and it, the computer, decides he should be told.
10	Computer does the whole job if it decides it should be done, and if so tells human, if it decides he should be told.

The Sheridan and Verplank model introduces a *computer* and a *human*, where the computer always *implements* the job. The basic idea of this scale is that the human to a varying degree will either tell the computer what to do (level 1), receive options from the computer before determining an action (level 2-6), or being told by the computer what the computer has already done (levels 7-10).

From a transportation perspective, the Sheridan and Verplank model has some challenges regarding *timing*. More specifically, several of the states cannot be regarded feasible in a real-time transportation scenario, where decisions must be made continuously.

2.2. SAE International's levels of driving automation

SAE International has developed a comprehensive *recommended practice,* SAE J3016 [6], providing what is called "taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles". SAE J3016 defines five levels of automation, excluding a zeroth level of no automation, ranging from "Driver Assistance" (level 1) to "Full Driving Automation" (level 5). It also introduces the following concepts, being fundamental to the understanding the levels summarized in table 2:

- **Driver**. A human user of a vehicle.
- **System**. A driving automation system that can operate the vehicle.
- **Dynamic Driving Task (DDT)**. All operations required to operate a vehicle in on-road traffic.
- **Object and Event Detection and Response (OEDR)**. The subtasks of DDT that include monitoring the driving environment and executing appropriate response.
- **Operating Design Domain (ODD)**. Operating conditions under which a given feature is specifically designed to function.
- **DDT Fallback**. The response by the Driver or the System after a performance-relevant failure or exit of ODD.

Table 2. Summary of levels of driving automation as defined in [5]

Level	Name	DDT: Sustained control of vehicle	DDT: OEDR	DDT Fallback	ODD
0	No Driving Automation	Driver	Driver	Driver	**n/a**
1	Driver Assistance	Driver and System	Driver	Driver	Limited
2	Partial Driving Automation	**System**	Driver	Driver	Limited
3	Conditional Driving Automation	**System**	**System**	Fallback-ready user[1]	Limited
4	High Driving Automation	**System**	**System**	**System**	Limited
5	Full Driving automation	**System**	**System**	**System**	**Unlimited**

[1] User becomes the driver during fallback

The SAE levels can briefly be summarized as follows:

- In level 1 the "Sustained control of vehicle" is split between the Driver and the System. This means that the System provides either steering OR acceleration/brake support to support the Driver.
- In level 2 the "Sustained control of vehicle" is held by the System, but the driver is responsible for Object and Event Detection and Response (OEDR) and handling any unforeseen situations (DDT Fallback).
- In level 3 the System handles the complete "Sustained control of vehicle" and OEDR, but the Driver must be ready to handle any unforeseen situations (DDT Fallback).
- In level 4, the System handles all foreseen and unforeseen situations, but can only operate within a limited Operating Design Domain (ODD).
- In level 5, the System handles all foreseen and unforeseen situations, and can operate under any Operating Design Domain (ODD).

Special attention should be drawn to level 3, where the System monitors the situation while the Driver still must be ready to become the driver during fallback, at short notice. This means that the Driver is ultimately responsible for the vehicle, and that s/he may have to handle a situation where the System has put the vehicle in immediate danger. Inagaki and Sheridan [11] suggest solving this problem by revising level 3 or by introducing haptic shared control, and they state that there is a need for more knowledge on how to trade authority when there is a *request to intervene*. However, one should also consider the possibility that the SAE levels have a basic logical flaw that makes it difficult to properly integrate the four main characteristics[2] that the SAE levels are built upon.

2.3. NFAS' levels of autonomy for merchant ships

The Norwegian Forum for Autonomous Ships (NFAS) has suggested four levels for operational autonomy for merchant ships [7], as described in table 3. The four levels have similarities with both [2] and [6], and the model also introduces a remote Shore Control Centre that may supervise or control the ship along with personnel on the bridge.

Table 3. Autonomy levels as suggested by the Norwegian Forum for Autonomous ships [7]

Name	Description (abbreviated)
Decision support	The crew is in direct command of ship operation and continuously supervises all operations. This level normally corresponds to "no autonomy".
Automatic	The operation follows a pre-programmed sequence and will request human intervention from either Shore Control Centre or the bridge if any unexpected events occur, or when the operation completes.
Constrained autonomous	The ship can operate fully automatic in most situations and will call on human operators to intervene if problems cannot be solved within defined constraints. Shore Control Centre or bridge personnel continuously supervises the operation and will take immediate control when requested by the system.
Fully autonomous	The ship handles all situations by itself and will not have a Shore Control Centre or any bridge personnel at all.

[2] Sustained control of vehicle (DDT), OEDR (DDT), DDT Fallback and ODD

The differences between the levels called *Automatic* and *Constrained autonomous* seem mainly to relate to the dynamic properties of the ship systems, and whether these follow a *pre-programmed sequence* or operate *fully automatic*. Since distinguishing a *pre-programmed sequence* from a *fully automatic* algorithm is not an exact science, this may indicate that two middle levels have some overlap. Furthermore, the *Constrained autonomous* level does not present any information on how to handle a situation where communication with Shore Control Centre is lost while no personnel is available on board. Without going into any further analysis, these two findings may indicate that the autonomy levels suggested by NFAS are partially overlapping and not fully complete.

3. Defining autonomy

Current approaches to describing levels of automation seem to have three common characteristics: (1) they aim to describe and categorize technological capabilities and functions, (2) the models tend to have flaws, limitations and level overlap, and (3) when fixing the models, new problems arise. One explanation for this is that we simply have not yet found an adequate technical categorization for autonomy, and that we have to dive further into the details. Another possible explanation, reflecting a harsh truth of science, is that even though some models intuitively feel right, they may actually be wrong.

Traditionally, the question of autonomy has been related to defining the capabilities a system must have to be considered *autonomous*. There is however an even more open question that could be raised: *When is a system considered autonomous*? If we accept that the defining characteristic of an *autonomous system* is its ability to relieve humans of responsibility by *assuming accountability* for an operation, the question of autonomy can potentially be regarded a legal question instead of a technical one. Reformulating this characteristic may then give us the following definition of autonomy:

A system is considered *autonomous* if it can *legally accept accountability* for an operation, thereby assuming the accountability that was previously held by either a *human operator* or another *autonomous system*.

Three aspects of this definition deserve specific attention:

(1) Regarding ***legally***. If autonomy is regarded a legal question rather than a technical one, any transfer of accountability from a *human operator* to an *autonomous system* must be pre-approved by relevant authorities.
(2) Regarding ***accept***. If an autonomous system should be able to accept accountability for an operation, it must also be allowed to *decline assuming accountability* for an operation. Without this option, the system in question would simply be a direct-controlled component in a larger system.
(3) Regarding ***accountability***. By the above definition, an autonomous system can relieve a *human operator* of being accountable for an operation. As only humans can be held legally accountable in current legislation, this means that the legal accountability for the operation is ultimately transferred to the creator of the system, the *system designer*.

Using the suggested definition for autonomy as basis, we can categorize any technical system into one of the two following groups:

- **Autonomous system**. A system that can legally accept accountability for an operation.
- **Non-autonomous system**. A system that cannot legally accept accountability for an operation.

Dividing these categories into even more sub-categories (e.g. separating non-autonomous systems into automatic and manual) is alluring, but neither can nor should be performed on the basis of the

suggested definition. Therefore, we will stick strictly to the above two levels of autonomy for the remainder of our analysis.

An interesting aspect with the above definition of autonomy is that is also holds when the *system in question* is a human. With regard to the operation "driving a car" a person holding a driver's license can be regarded as autonomous, as s/he can legally accept accountability for driving the car. A child will on the other hand be regarded as non-autonomous, as it cannot legally accept accountability for using the car. Furthermore, a person might in this respect be "autonomous" in one country, but not in another. Such geographic differences should be expected for "non-human" autonomous systems as well, with different countries having different laws regarding certification and use of autonomous systems.

4. Categorizing systems as autonomous or non-autonomous

When analyzing whether a system is autonomous or not, one first needs to specify the roles of the *operator* and the *system designer;* the *operator* is the one using the system, while the *system designer* has created the system. Second, one must ask who will be held accountable for any incidents caused by the system; the operator or the system designer. Note that if a system accepts accountability for an operation, the system designer will ultimately be accountable for any incidents caused by the system. **Therefore, autonomy is inextricably related to the exchange of accountability from an *operator* to the *system designer*.**

4.1. Example 1: Elevator

Scenario: A person is using a standard elevator in an office building or hotel.

There are three roles to consider in this case:

- **The system:** The elevator
- **The operator:** The elevator user
- **The system designer:** The elevator designer (i.e. the person being technically responsible[3] for the elevator)

Questions of accountability:

- Is the *elevator user* accountable for any incidents caused by him/her calling the *elevator*? **No**
- Is the *elevator designer* accountable for any incidents caused by the *elevator user* calling the *elevator*? **Yes, because the deployment and operation of elevators is legally regulated.**

In this example the elevator user (i.e. the operator) is not accountable for any incidents related to the elevator (i.e. the system), meaning that **the elevator should be considered autonomous**.

4.2. Example 2: Remote-operated unmanned surface vessel (USV)

Scenario: A person performs remote-control of an unmanned surface vessel (USV).

There are three roles to consider in this case:

- **The system:** The USV
- **The operator:** The USV operator, who remotely controls the USV
- **The designer:** The USV designer (i.e. the person being technically responsible for the USV)

[3] Note that being responsible for a system does not necessarily mean that one is accountable for the system's actions. A car mechanic is responsible for fixing your car but is not accountable for your driving.

Questions of accountability:

- Is the *USV operator* accountable for any incidents caused by him/her deploying and operating the *USV*? **Yes, as there are currently no laws allowing a USV to assume accountability for its operation.**
- Is the *USV designer* accountable for any incidents caused by the *USV operator* deploying and activating the *USV*? **No, the USV Operator cannot transfer accountability to the USV.**

In this example the USV operator (i.e. the operator) is accountable for any incidents related to the USV (i.e. the system), meaning that **the unmanned USV should <u>not</u> be considered autonomous**.

4.3. Example 3: Unmanned metro train

Scenario: A person performs continuous remote supervision of an unmanned metro train over video link, and is required to activate a remote-operated emergency stop in case of emergency.

There are three roles to consider in this case:

- **The system:** The metro train
- **The operator:** The metro train supervisor
- **The designer:** The metro train designer (i.e. the person being technically responsible for the metro)

Questions of accountability:

- Is the *metro train supervisor* accountable for any incidents caused by him/her activating the *metro*? **Yes, because s/he is required to activate remote-operated emergency stop in case of emergency.**
- Is the *metro train designer* accountable for any incidents caused by the *remote metro supervisor* activating the *metro*? **No, because the metro train supervisor is required to continuously monitor the situation.**

In this example the metro train supervisor (i.e. the operator) is accountable for any incidents related to the metro train (i.e. the system), meaning that **the unmanned metro train should <u>not</u> be considered autonomous.**

5. Integrating responsibility-based autonomy into existing frameworks for autonomous ships

The NFAS Definitions for Autonomous Ships [7] suggest that the three states *Remote control, Automatic ship* and *Constrained autonomous* (see table 3) can be used in combination with fully or periodically unmanned bridge. However, these three states translate into a non-autonomous system according to the definition suggested in this paper, meaning that there would be no *accountable system or human* on board in these cases. As communication links to the Shore Command Centre cannot be made 100 percent reliable, this may potentially indicate that *autonomy is a prerequisite for unmanned remote-controlled operations* in order to maintain on-board accountability. This is in contrast to the common view that remote-controlled operations are more easily achieved than autonomous operations, and may suggest that the principles behind [7] need to be reconsidered.

To our knowledge, there are not yet any formal international framework for commercial operation of autonomous ships. However, if such a framework should be established, we expect that it would be rooted in existing IMO regulations such as SOLAS [12], COLREG [13] and STCW [14]. Going forward in a revision of these regulations without a common definition and understanding of the concept of autonomy seems very unfortunate given the inconsistencies presented in this paper. We therefore strongly advise against passing regulations for autonomous operations without first defining what is actually meant by *autonomy*. The definition suggested in this paper is an attempt at paving the way towards such a common understanding, but we do neither hope nor expect that any suggested ideas will remain unchallenged by the industrial and scientific community. We do however hope that

our analysis has illustrated that the concept of *autonomy* should be formally defined, and that technical capabilities are not necessarily the only foundation for such a definition.

6. Conclusions

This paper has suggested that defining autonomy as a technical capability may not be feasible, and proposes a definition of autonomy based on legal accountability rather than on functional capabilities. The background for this idea stems from the experience that existing descriptions of levels of automation contain ambiguities and are partially overlapping, and that they do not give any concise definition of autonomy. After having suggested a new definition for autonomy, the paper demonstrates how autonomy can be understood in the context of real cases. This includes illustrating how autonomy could be regarded as a question of certification and regulation more than of technical complexity. Finally, the paper points to potential challenges related to the existing definitions for autonomous ships and proposes looking into these with the suggested definition of autonomy in mind.

As this paper has merely scratched the surface of autonomy from the perspective of responsibility and accountability, several questions are still to be answered. One topic is how accountability should be transferred between humans and autonomous systems, both practically and formally. Another question is how accountability should be handled or split between sub-systems from a wide range of suppliers. Also, one might need to consider investigating split accountability between an operator and a system, and how this potentially differs depending on whether the operator or the system holds the initial overall accountability.

Acknowledgments

The authors acknowledge the support of the H2020 ECSEL SafeCOP project (Grant 692529-2), the H2020 GSA H2H project (Grant 775998), and the Sacomas project funded by the Research Council of Norway (Grant 269510).

References

[1] National Research Council (2005) Autonomous Vehicles in Support of Naval Operations. Washington, DC: The National Academies Press. https://doi.org/10.17226/11379
[2] Sheridan T B and Verplank W L (1978) Human and computer control of undersea teleoperators (Department of Mechanical Engineering, MIT, Cambridge, MA, USA)
[3] Endsley M R (1987) The Application of Human Factors to the Development of Expert Systems for Advanced Cockpits. Proceedings of the Human Factors Society Annual Meeting, 31(12), 1388–1392. https://doi.org/10.1177/154193128703101219
[4] Endsley M R and Kaber D B (1999) Level of automation effects on performance, situation wareness and workload in a dynamic control task. Ergonomics, 42(3),462-492, https://doi.org/10.1080/001401399185595
[5] BASt (2013) Legal consequences of an increase in vehicle automation: final report of the project group. BASt-Report F83
[6] SAE International (2018) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (Surface Vehicle Recommended Practice: Superseding J3016 Sept 2016)
[7] Norwegian Forum for Autonomous Ships (2017) Definitions for Autonomous Merchant Ships. http://nfas.autonomous-ship.org/resources/autonom-defs.pdf (Retrieved 2019-02-28)
[8] U.S. Department of Transportation (2018) Preparing for the Future of Transportation: Automated Vehicle 3.0. https://www.transportation.gov/av/3 (Retrieved 2019-02-28)
[9] https://sae-europe.org/2018-global-ground-vehicle-standards-summary-year-advancing-sae-ground-vehicle-standards-global-transformed-mobility-arena/ (Retrieved 2019-06-15)
[10] National Research Council (1998) The Future of Air Traffic Control: Human Operators and automation. Washington, DC: The National Academies Press. https://doi.org/10.17226/6018
[11] Inagaki T and Sheridan T B (2018) A critique of the SAE conditional driving automation definition, and analyses of options for improvement. Cognition, Technology & Work, 1-10

[12] International Maritime Organization (1974) International Convention for the Safety of Life at Sea (SOLAS)
[13] International Maritime Organization (1972) International Regulations for Preventing Collisions at Sea (COLREG)
[14] International Maritime Organization (1978) International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW)