Holistic bow-tie model of meaningful human control over effective systems

Towards a dynamic balance of humans and AI-based systems within our global society and environment

Flemisch, Frank; Baltzer, Marcel; Abbink, David; Cavalcante Siebert, L.; Diggelen, Jurriaan van; Herzberger, Nicolas Daniel; Draper, Mark; Boardman, Michael; Pacaux-Lemoine, Marie Pierre; Wasser, Joscha

DOI
10.4337/9781802204131.00025

Publication date
2024

Document Version
Accepted author manuscript

Published in
Research Handbook on Meaningful Human Control of Artificial Intelligence Systems

Holistic Bow-Tie Model of Meaningful Human Control over Effective Systems: Towards a dynamic balance of humans and AI-based systems within our global society and environment

Frank Flemisch, Marcel Baltzer, David Abbink, Luciano Cavalcante Siebert, Jurriaan van Diggelen, Nicolas Daniel Herzberger, Mark Draper, Michael Boardman, Marie-Pierre Pacaux-Lemoine, Joscha Wasser

While Meaningful Human Control (MHC) is at the very heart of the Edward Elgar research handbook, this specific chapter addresses the questions how MHC is rooted in the history of human artefacts and human-machine systems, how it is related to the term control, ability, responsibility, authority, autonomy and finally accountability. The chapter sketches, step by step, a holistic, cybernetic model of the most important relationships between MHC and its related concepts interconnected over this holistic "big picture" map.
Starting point are existing control systems and their evolution through history, followed by the interrelationship between the small-scale human-machine or human-AI system, and the increasingly bigger system of systems, organizations, societies and our global environment. The goal of this bow-tie shaped system map is to enable a better balance between global and local perspectives, and therefore enable a more efficient and better design, engineering and evaluation of such systems.

Keywords: Holistic, Bowtie diagram, cybernetic, control, controllability, Meaningful Human Control

Introduction

With increasing intelligence and more and more autonomous abilities of technical systems and their combination within systems of systems, it becomes increasingly important that usability, transparency and controllability of such systems are at the center of every research, design and development activity.

While Meaningful Human Control (MHC) is at the very heart of the Edward Elgar research handbook, this specific chapter addresses the questions how MHC is rooted in the history of human artefacts and human-machine systems, how it is related to the term control, to effectiveness, and to other systemic concepts and attributes, especially ability, responsibility, authority, autonomy and finally accountability. The ambition of this chapter is to sketch, step by step, a holistic, cybernetic model of the most important relationships between MHC at the very center and its related concepts interconnected over this holistic "big picture" map.
Starting point will be the analysis of existing control systems and its evolution through history. Special attention will be given to the interrelationship between the small-scale human-machine or human-AI system, and the increasingly bigger system of systems, organizations, societies and our global environment of planet Earth, in which it is embedded. The goal of this bow-tie shaped system map is to enable a better balance between the different perspectives, and therefore enable a more efficient and better design, engineering and evaluation of such systems.

A brief history of human control and human systems integration

In order to shape the future of humans and human artifacts like AI-based Systems, it can be useful to understand the past of humans and human artifacts. The chapter therefore starts with a historic perspective on intelligence and control initially of humans, other animals and nature, and increasingly of human artifacts like tools and machines.

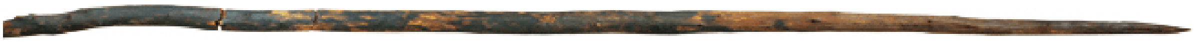


Figure 1: Spears of Schoeningen as an early example of Meaningful Human Control

Figure 2: Man wounded or killed by several spears, painted in red ochre. (c.20,000 BCE; Pech-Merle-Cave): Effective or meaningful Human Control, or loss of control? (visual-arts-cork.com, 2022)

Figure 1 shows one of the throwing spears of Schöningen, which were excavated between 1994 and 1999 in an open-cast coal mine in the north of Germany, and are dated around 300.000 years back, making them the oldest complete wooden weapon found on Earth so far (Thieme, 1997). These spears are an early example of Human Factors, for which Homo Heidelbergensis, a pre-runner of Homo Sapiens, was already able to combine different techniques like cutting and fire to carve an effective tool and adapt it to the individual bearer. These spears are also an early example of Human Systems Integration, for which not only the optimization of the tool itself is important, but also the integration into an effective organization and cooperation of several individuals, here Homo Heidelbergensis, hunting together so effectively that even larger or faster prey like elephants or horses could be hunted. Tomasello (2014) describes how human cognition evolved together with the ability to create and handle such tools, and especially how the cooperation and shared intentionality fostered the evolution of Homo towards Homo Sapiens as one of the most dominant species on this planet. Later, Coss (2017) show that the same visuomotor coordination that enabled Homo sapiens' ability to accurately throw a hunting spear went hand in hand with their ability to create realistic cave drawings.

Flemisch et al. (2019b) sketch how these spears can be used as a metaphor to describe shared and cooperative control between humans and machines. Flemisch, Preutenborbeck et al. (2022) use the spear as an early example of disruptive technology, which changed the balance of power dramatically, as also the example in Figure 2 shows.

For us, the spears of Schoeningen are also an early example of human control over an artifact: It must have been increasingly essential for the bearers of those spears to control when and where this spear was flying and which target it was hitting. It is also one of the earliest examples of a tool not only for which control plays an important role, but moreover, which may get out of hand and becomes uncontrolled, i.e. cannot longer be influenced strong enough for a time period before it becomes effective.

Even this early example, can already be used to discuss meaningful and control and its relationship to effectiveness, more about this later. Other, later examples are bow and arrow, which have been used e.g. by Miller (2022) to describe Ethical Neglect Tolerance.

While the spear and its interaction with the environment and the target has already some complexity, which is difficult to control, it has no intelligence of its own, at least unless we believe in magic or Gods influencing the flight of a weapon like in Homer's epic poems. With guided weapons this has already changed, and will change even more by integrating more AI into artefacts.
Looking beyond spear-like artifacts at control of an intelligent entity, we also have a couple of historic examples. The most striking one is the interrelationship of Homo Sapiens and horses described as Equestrianism (from Latin equester, Horseman, equus horse), e.g. in horseback riding or horse cart driving.

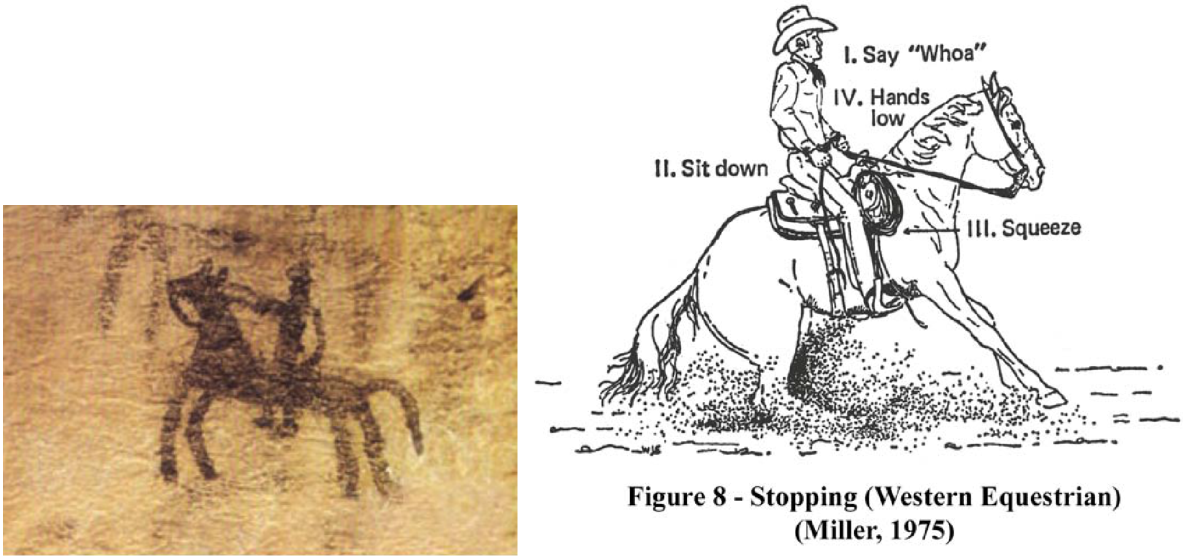


Figure 3: First highly automated, cooperative transportation (and defense) system

There might be some reasons that with a history of about 5,000 years (Chamberlin, 2010) horseback riding and cart driving are historically much younger than the use of spears. A reason might have been the different way to control. Obviously, a more complex way of control and controllability was necessary and had to evolve over thousands of years to really make use of horses. It has some reasons why humans don't ride tigers, and even with on the first glance similar animals like zebras it is much more difficult to tame them (Diamond, 1999). Some intelligent systems provide more ability for cooperation than others.

The most obvious key to control horses is an appropriate interface, here horse bits and reins, which interface the movement intent between the human and the animal and even allow a negotiation of direction, speed or complex maneuvers. Beyond the human-animal-interface is a whole cascade of other key concepts, including mutual understanding, training, trust building, arbitrating, cooperation and even symbiosis. These complex aspects were increasingly understood and appreciated, initially as a practical craftsmanship, then also scientifically, even by societies which do not use horses for food, economic or military reasons anymore (Chamberlin, 2006; e.g. Wanless, 2001, 1992)

The relationship between humans and horses can also be used as a design metaphor, i.e. as blueprint for the relationship between humans and intelligent non-human entities, e.g. as H-Metaphor initially described by Flemisch et al. (2003) and exemplified as H-Mode by Goodrich et al. (2006) or Altendorf et al. (2016), or - when it is realized as mutually adaptive interaction - as symbiotic driving (Abbink et al., 2018; Melman et al., 2020).

Ironically, the example of rider and horse and more generally of humans and their animals had more impact on the controllability of such automated cars than most people realize. One aspect is that the levels of vehicle automation e.g. described as levels of driving automation by SAE, based on Gasser et al. (2012) was based on an automation scale model derived from the H-Metaphor (Flemisch et al., 2008). Another aspect concerns legal issues, as Article 5 of the 1968 Vienna Convention on Road Traffic stated "5. Every driver shall at all times be able to **control** his vehicle or to **guide** his animals.", which opened up a whole loophole for expanding the controllability paradigm beyond the direct controllability on the control level of driving e.g. with a steering wheel(Hammond, 2015). More details on this further down.

Regarding the topic of this book, control, controllability and especially Meaningful Human Control, the metaphor of human and horse contains a couple of insights, e.g. that:

- Control is related to control loops, which can be actually built up and tested,
- Control can be traded or shared, and
- Control can be realized on different levels or layers, and at different levels of automation and autonomy and different layers of cooperation (Abbink et al., 2018; see e.g. Flemisch et al., 2019a; Pacaux-Lemoine & Flemisch, 2019)
-

Another example for the relationship of humans and intelligent entities is the relationship with dogs, which was already hinted by Hancock's insightful article on teleology for (intelligent) technology (Hancock, 1996), and also contains rich examples for controllability. With this example, it also becomes clear that control can be related to effectiveness, but also to other aspects like social, ethical or societal aspects, which are related here as meaningfulness. Effectiveness is not always directly related to meaningfulness, as the example of a drug dealer, effectively controlling his attack dog. In this article, we promote Meaningful Human Control over effective systems. An example for this would be a well selected and trained police dog, being effective on the one hand, but also under meaningful control of its police officer.
Systems with dogs or horses or other animals are designed regarding not only the mechanical or physical aspects, but especially considering the cognitive aspects of both the humans and the other animals in the systems. Animals were trained to be as deterministic as possible, but it was always clear that this determinism is per se limited, and that controllability, especially of unforeseen behavior is always an issue.
Ironically, with the rise of more mechanized artifacts in the late 19th and early 20th century, the focus of the engineers shifted from the cognitive towards the mechanical and physical aspects, with the goal to design these systems as deterministic (and thus predictable) as possible, and with a decreasing willingness and ability to deal with non-deterministic issues.

This trend changed again with the rise of the **computer**, starting with the first process controlled computer Z1 of Konrad Zuse in 1936 in Berlin, followed by the Mark I built 1944 by Howard Aiken and team from IBM, which was used to support the Manhattan Project. In close relation with this revolution in hardware, another scientific revolution happened regarding the software and scientific models: in 1948 Norbert Wiener wrote his famous book introducing Cybernetics, one of the most influential books of the twentieth century (Wiener, 1950, 2019). Interestingly, Wiener is often quoted to have written about computers, but his book is more general about control and communication mechanisms both in animals and machines, and also sparked system science beyond computers, especially in biology and psychology. "Control" is even in the title of his book, and as the core of cybernetics, he describes mechanisms of these systems to **self-control** themselves and the relationship to their environment. A core idea of cybernetics are feedback loops, which support to regulate and control. The new science of cybernetics was extensively discussed during the Macy conferences, between 1946 and 1953, by an interdisciplinary community of scholars and researchers from different disciplinary perspectives such as mathematics, psychology, anthropology, sociology, engineering, and economics (Pias, 2003/2016). These disciplines joined to lay the groundwork for the new science of cybernetics.

Since the rough start of cybernetics and control theory in the 1950ies, with control and controllability at its very center from early on, this section of science has gone a long way. It has seen several hype cycles as proposed by Gartner (e.g. Fenn & Time, 2007) , e.g. towards Knowledge Based Systems, later to Cognitive Systems, Neural Networks and, with the invention of deep Neural Networks, a big re-surge in the 2020s. However, unlike in the natural sciences, disciplines that worked towards these developments such as engineering, computer science, but also business, politics, architecture, and even painting are concerned not necessarily with how things are but how they might (or ought to) be – in short, with design (Simon, 1996/2019). Questions of control in automated systems can only be tackled by acknowledging ahead of time that one's ability to control is not a matter of chance. It is something that is strongly influenced by the design of the system and its socio-technical environment (Cavalcante Siebert et al., 2022; Siebert & Abbink, 2022) .
With increasing capabilities of controllers and AI and their use in a couple of domains like transportation, aviation or defense, the demand for transparency (e.g. Chen et al., 2020) and controllability (e.g. Flemisch et al., 2012) of those systems, already mentioned by Wiener (1950, 1954, 2019), surfaced again, and led to the discussion of MHC.

Figure 4: Future Combat Air System FCAS as an example for a 21st century highly automated, cooperative defense system, with a mix of inhabited and uninhabited highly automated vehicles (Airbus 2021)

The starting point for controllability issues was actually in the **aviation** domain. One of the first mentions of control was the Cooper-Harper Rating Scale of Handling Qualities, which has the question "Is it controllable?" at the very heart of the 10 point scale (Cooper & Harper, 1969). Since then, a gradual revolution towards highly automated, often AI based systems took place. The first human factors problems, many related to system transparency and controllability, were discovered (e.g. Billings, 1996), the disciplines of Human Factors and Human Systems Integration founded (e.g. by NASA and US Department of Defence DOD), and cognitive assistant systems explored (for an overview see e.g. Onken and Schulte (2010). Starting with experimental systems in the 1990ies, UAV (Uninhabited Air Vehicles), later called UAS (Uninhabited Air Systems) were developed and implemented by the US Airforce first, later operated by almost any military in the world. Controllability, and especially resilience against loss of control, is at the very core of these systems.

In the 2020s, intense research and development was directed towards the integration of inhabited and uninhabited systems, e.g. the French-German FCAS (Future Combat Air System, Figure 4). Controllability of such powerful systems is, not surprisingly, at the very core of the discussion (Figure 4, FCAS 2022).

Another striking **example** for controllability issues is the **automotive domain**: Driving automation is possible due to the direct application of cybernetics and control theory, combining functionalities of several automation systems. One of the first steps for modern driving automation was the development of the first cruise control by Ralph Teetor in 1945 (Teetor Meyer, 2011). Additionally to such basic automation functionality of keeping on a fixed trajectory, future automated vehicles need the ability to "see" their boundaries, such as road markings, traffic rules etc.. In order to achieve that, technology such as laser scanners (LIDAR) or high resolution cameras is needed in combination with computer vision algorithms, which in turn require extensive computing power.

After a start as niche research with the presentation in 1977 of the first "autonomously driving car" at Tsukuba Mechanical Engineering Lab in Japan (e.g. Tsugawa, 1994) and milestones like the European "Prometheus" project (e.g. Dickmanns, 1998) and the DARPA Grand Challenges (e.g. Thrun, 2006), in 2022, most automotive manufacturers research and develop automated vehicles.

One of the most active companies in this field is Waymo (originally Google Driverless Cars) having been granted the first testing license for the US state Nevada in 2012 (Warren, 2012). Tesla was one of the first companies to sell vehicles with a function termed "Autopilot", which however was a partial driving automation (SAE Level 2) that still requires the driver's full attention throughout the journey. Flemisch et al. (2017; 2016) described an uncanny and unsafe valley of automation with controllability problems between an SAE-level 2 and an SAE level 3 vehicle and used the Tesla "Autopilot" as an example of a system right in the unsafe valley. Shortly after that, a Tesla Model S was also involved in the first fatal crash of an automated vehicle where the driver died while using the "Autopilot". The first crash where a pedestrian was killed was caused in 2018 by a highly automated test vehicle with safety driver, operated by Uber, with operators complacency and controllability being at the very core of the accident investigation (National Transportation Safety Board, 2019).

Despite the problems with some level 2 systems, the first homologated Conditional Driving Automation (SAE level 3) was granted on December 2nd 2021 to Mercedes for their "Drive Pilot", a highway traffic assistant system which can be purchased as part of a 2022 S-class or EQS.

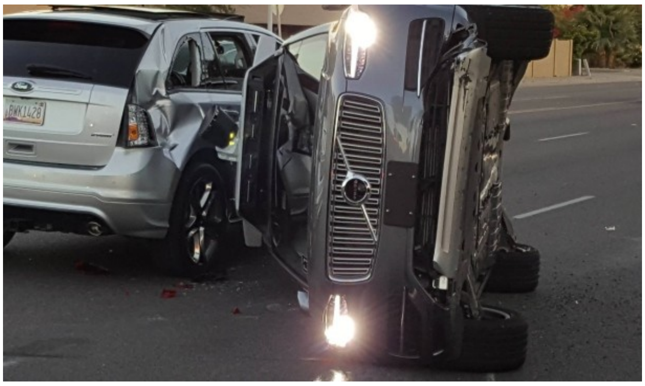


Figure 5: 21st century highly automated, cooperative transportation system (here a situation after of loss of control)

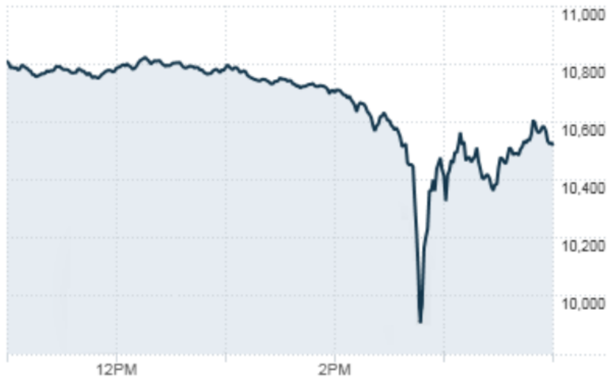


Figure 6: Stock market and flash crash as an example for insufficient control, leading to instabilities

Another **example** of a complex system spinning out of control appeared in highly automated **stock trading**. In 2018, approximately three-quarters of all trades made in the U.S. stock markets are done by bots (Scharre, 2018). These bots are not bound by any physical speed limitations: they act in the information space and can make decisions faster than any human. Furthermore, once a stock trading bot proves useful for a user, they can be multiplied infinitely and perform work for their users. This is where potential control problems appeared: the speed and the scale at which they operate is no longer manageable by humans. On May 6, 2010, a large amount of stock trading bots worldwide (together with other factors) caused a flash crash on the stock market (Figure 6). After this flash crash, extra regulations were introduced on the stock market to deal with these new phenomena.

Trading at machine speed is already more commonly used, **fighting at machine speed** has already been used since the 1960s in aviation, air defense and the nuclear forces, and now being developed and rolled out to all military systems (e.g. Scharre, 2018). The advantage of the machine gets even bigger when processes are completely running in cyber space. Concerns regarding the possible deployment and use of weapons with autonomous capabilities have led to a joint effort on determining what should be considered acceptable when using armed forces with such weapons. In this context, the term Meaningful Human Control was coined (Boillot, 2014), regarding the exercise of control over the use of weapons and the related responsibility and accountability for its consequences. With new, faster weapon systems like hypersonic missiles, and especially with the interconnection with internet based warfare (cyberwar), it becomes increasingly clear why controllability of such systems is in the focus of a couple of working groups e.g. at NATO (e.g. Boardman & Butcher, 2019; Draper & van Diggelen, 2020)

Step by step from essence of human control towards a holistic model

How are all these complex examples connected? As for the understanding and shaping of complex systems, there is “nothing as practical as a good theory” (Lewin, 1943). The chapter is based on system theory, and continues with a step-by-step buildup of definitions, descriptions, a system model and a system of systems models of MHC and their related concepts. As theory would be useless without practice, each of the theory blocks is enriched with system engineering and human systems integration research and practice, with examples from real systems e.g. from aviation or the automotive domain.

Step 1: Influence and control as an essential part of (human) life

Step 1 addresses one of the main challenges of modeling: What aspects and their relationships of a complex reality, here about control, are so essential and so general, that they should be at the very center of the model? What is an appropriate abstraction or simplification level of these aspects that a) helps to understand real issues (pragmatic argument) and b) is backed by a theory that increases the likelihood that it is true (theory argument).

In theory, control is generally being understood as

a)"the ability or power to decide or strongly influence the particular way in which something will happen or someone will behave [] (Cambridge Dictionary, 2014) or
b)" to exercise restraining or directing influence over,
c) to have power over,
d) to reduce the incidence or severity of especially to innocuous levels", e.g. of insects (Merriam-Webster).

Applied to human-machine and human-AI systems, for us **to control a system or a situation means to have enough influence on the situation that it develops or keeps in a way or region preferred by the controlling agent (see also Flemisch et al., 2012)** (see also Flemisch et al. 2011). For us, to control is related to influence, but stronger, with an impact beyond a threshold.

As Wiener (1950) already sketched, the fundamental mechanism in biology and technology for influencing or controlling a part of the world are feedback loops. Strongly oversimplified (Figure 7), agents perceive information from the environment and, based on prior knowledge, use this information to act on the situation in such a way that the situation develops towards preferred or "good" situations, and to act non-preferred or "bad" situations. Later on in Control Engineering, one of the many disciplines sparked from Cybernetics, this was described as closed loop control, versus open-loop control where no feedback is needed from the situation. Control Engineering also differentiates between continuous closed-loop control and discrete closed-loop control, where a triggering event is needed before the loop is closed again. In a later version of his book, Wiener (1961) described learning machines, which was later on realized e.g. as model based controller. Here, the perception of the situation is not only used to act directly on it, but also to adapt an internal model of the situation and the control itself in a way that the control becomes ever more efficient and effective (e.g. van den Hof et al., 2009).

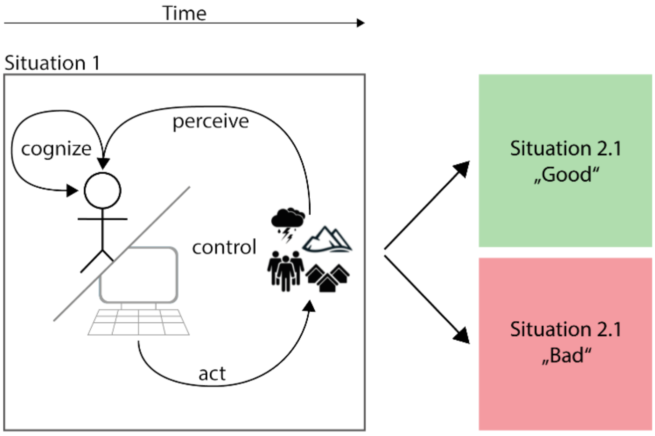


Figure 7: Human OR Computer, and process ("plant") form a system with feedback and control loops, to achieve good and avoid bad situations (inspired by Wiener (1950) and Franklin et al. (2002))

It is important to note that the simplicity of the model is per design, so that we have a starting point for the discussion to be as simple and clear as possible. The world itself is obviously much more complex and interrelated. However, even with this simple starting point, we can already discuss a couple of fundamental issues:

- **What is good or bad is not an absolute, but a subjective concept**. It strongly depends on the controlling agent and its reference values. In an expansion of the model, the controlling agent is just one of several stakeholders, who might have different reference values of what is good or bad.
- The **discrete situation space** described here, either one “good” and one “bad” situation **is the strongest oversimplification**. In more sophisticated models about reality, this could be more a continuum of situations judged with a spectrum of greyscales, or all colors of the rainbow especially when more than one quality is important to evaluate and judge a specific situation in a reasonable manner. This can be shown by an example from Cognitive Warfare, where “good” and “bad” is differentiated, using the essence of signal detection theory, into “correct action”, “incorrect action”, “correct non-action” and “incorrect non-action” (Flemisch, 2022).
- **In order to control the situation, the agent needs a minimum of ability, here of perception, action, and, with increasing complexity, also of cognition.** In the philosophical literature, this is usually known as the knowledge or epistemic condition for responsibility. For more complex and dynamic environments, this ability might also have to include the ability to reason or learn, i.e. to be intelligent.-
- **In order to control the situation, it becomes also clear that the agent needs a minimum of autonomy.** In the philosophical literature, this is known as the control or freedom condition (Talbert, 2019). If there are other entities influencing the situation too strongly, it does not make sense anymore to speak about control, but it can still make sense to talk about influence.

Our debate so far has been mainly from a cybernetic and control theoretic perspective, addressing more an **effective control**. With the concepts of ability and autonomy, it becomes increasingly clear that these perspectives would not be enough to describe our desire for control, and that humans and societies desire for not only effective, but more, which we call **meaningful control** of complex and safety critical systems. For that, it makes sense to take a philosophical perspective at first, before we return to a system perspective later:

The philosophical account of Meaningful Human Control proposed by Santoni de Sio and van den Hoven (2018) is based on the debate of free will and moral responsibility, mainly influenced by the concept of guidance control (Fischer & Ravizza, 2000). The concept of guidance control tries to reconcile **moral responsibility** with control, without the necessity of involving alternative possibilities. In other words, the concept of guidance control recognizes that one might be considered responsible for an undesired outcome even if the person could not do anything else at the time of the mishap, but might have done something before to avoid the undesired outcome. Consider the example of a drunk driver causing a road accident. Even though one might argue that the driver, because of the high level of intoxication, did neither have the minimum perception or cognition to steer the vehicle (knowledge condition), nor had the choice to do something else to avoid the accident at the time of the accident (freedom condition), the driver can still be considered blameworthy. The argument here is that the driver chose to drink knowing that they would need to drive later. A parallel can then be drawn between this case and an AI-based system with some degree of autonomy. In guidance control, the view shifts the focus from the agent to the mechanism or process leading to the action. Possession of guidance control (partially) depends on whether an agent’s mechanism is responsive to the controlling agent’s or the stakeholders’ situation and value space (e.g. “good” or “bad” situations).

Applied to real systems, an example for the described control loop in intelligent transportation is a **driver in a vehicle or a pilot in an airplane**, who perceived the environment, and based on perceptions and cognitions acts on the vehicle in a way that it keeps on driving towards the desired goal, and not causing an accident. Another example would be an autonomous vehicle, where an automation replaces the driver or pilot, and controls the vehicle in a way that a desired outcome is likely.

Step 2: Differentiating more between steps of guidance and control

Step 2 adds an additional level of complexity to the model and tries to simplify it in a form that a) the complexity of the model is still manageable, and b) it provides a connection to the higher complexity behind the model.

In theory, the fundamental model of perception, cognition and action already allows us to discuss the most fundamental steps in controlling a situation. In order to describe it even in more detail, over the years many more models were invented. As one out of many, the OODA loop originally described by Boyd (1996) for war situations, describes perceiving, thinking and acting of agents in four stages of

- **O**bservation: Gathering of outside information and matching them with unfolding circumstances and unfolding environmental interaction.
- **O**rientation: Judging the observation in the light of previous experiences, genetic heritage and cultural traditions.
- **D**ecision: Selecting one of several hypotheses, and put them to test.
- **A**ction: Implementing the decision

It is important to note that John Boyd was not a control engineer nor a cognitive scientist, but a former fighter pilot and military consultant with two bachelor’s degrees in economics and industrial engineering. Regardless of any academic ranks, his model, which is essentially a system dynamics model, contains quite some valuable insights, especially about implicit and explicit **guidance and control** feedforward and feedback loops. It especially describes how an essential loop of perceiving, thinking and acting, e.g. to control a situation, is supported by a cascade of feedforward and feedback loops, and especially how it is influenced by heritage and belief systems. In hindsight it is not clear how much of philosophical background John Boyd implemented, but the parallels to second-order cybernetics and radical constructivism are obvious, which describes how humans construct their reality based on a-priori-knowledge (Foerster & Poerksen, 2002; e.g. von Glasersfeld, 1984).

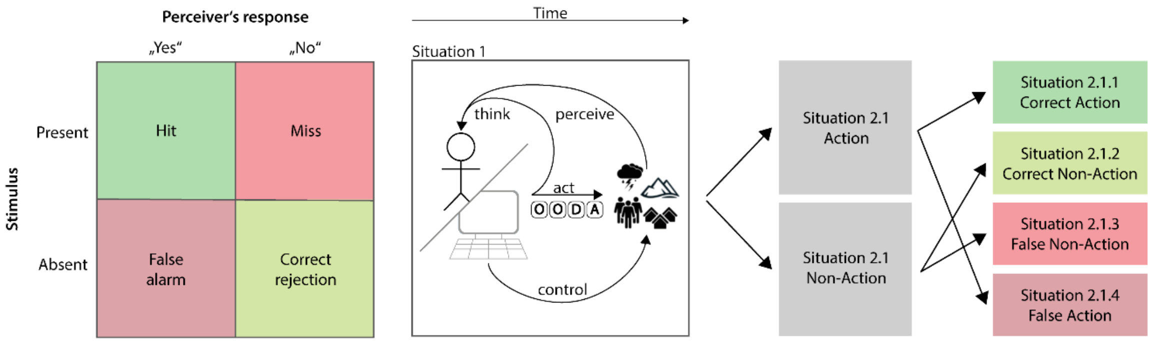


Figure 8 left: Signal detection theory
Figure 8 right: dilemma model for sensing and acting under uncertaintyHuman OR computer perceive, think, act and control, example OODA-Loop

Applied to our example domains, the OODA loop is being used to describe the control of pilots over fighter vehicles, but also for much more complex cycles like the targeting cycle of allied force actions (e.g. Daniels, 2021) or for cyber defense of armies or countries (e.g. Clarke & Knake, 2019).

Figure 8 also shows another important modification of our model, the refinement of what is good or bad over time. Often in time critical situations a decision for an action or non-action is needed, but whether this action leads to a good outcome, i.e. is regarded as good or correct action, can only be made in hindsight. This is inspired by and can be combined with the signal detection theory, where the decision of a perceiver to respond to a stimulus can either be a hit, where the stimulus is indeed present, or a false alarm, where the stimulus is in fact not present, or a miss, where the stimulus was present but not discovered, or a correct rejection, where a stimulus was not present and also not detected. A similar model can be applied to action, which might be based on a signal detection, but also on other stages of decision making e.g. described by the OODA-loop: A decision for action or non-action, and its implementation could in hindsight be judged as correct or false action, or in case of a non-action, a false or correct non-action.
Thinking in such a stage wise approach of action/non-action and judgement helps to understand the fundamental dilemmas of decision making and action especially under uncertainty, why we call this the dilemma model of sensing and acting under uncertainty.

Applied to defense systems, Flemisch (2022) describes an example in Cognitive Warfare, where a Tactical Control Officer decides on shooting down an airplane, based on the advice of an AI based advisory system.

Step 3: Joint perceiving, thinking and acting as part of joint cognitive systems

Step 3 takes one step forward, from the single agent to more than one agent controlling. What are the additional factors, which are not yet in the single agent model, and which are so important, that they should be included in a more holistic model, hopefully by keeping it simple enough that the model is still manageable?

In reality, often more than one agent takes part in influencing or controlling a situation. For humans, thinking and working together is quite natural. Already with the spears of Schoeningen, as described above, cooperating must have been a key concept for the successful use of these tools. Also Boyd's OODA-Loop (Boyd, 1996) is today mainly applied to the joint action of several individuals, even of several organizations or nations.

In theory, it took some time until theoretical frameworks on action caught up on the reality of joint action and cooperation. Only in the last decades, anthropologists started to understand also in detail the unique ability of Homo species to cooperate. Harcourt and Waal (1992) describe this ability for human and non-human species, but especially how Homo Sapiens excels in the complex cooperation with other species, and also with different species. Tomasello (2014) describes **shared intentionality** and the ability to cooperate towards common goals as the base of this cooperation.
Starting with the first concepts of psychology on how people interact, a research track on "**joint action**" was formed to describe "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment." (see e.g.

Sebanz et al., 2006). In parallel efforts, the neural basis of cooperation, mirror neurons, have been uncovered not long ago (e.g. Rizzolatti & Sinigaglia, 2008). There are strong hints that shared mental models form the base for joint reasoning and joint action.

Distributed cognition is another framework, as proposed by Hutchins (1995), describing cognition in terms of the emergence and interactions of component parts. The theory of distributed cognition focuses not on how individual actors make decisions considering social and environmental features but on a broader class of cognitive events that surpasses the individual. It is an approach to understanding cognition from a distributed perspective across members of a group, environment and through time.

Applied to our challenge of control, it becomes increasingly clear that this challenge of Meaningful Human Control is related to joint action and joint cognition as well, where shared mental models are needed in order to align or complement perception, cognition and action in a way that serves the common goals.

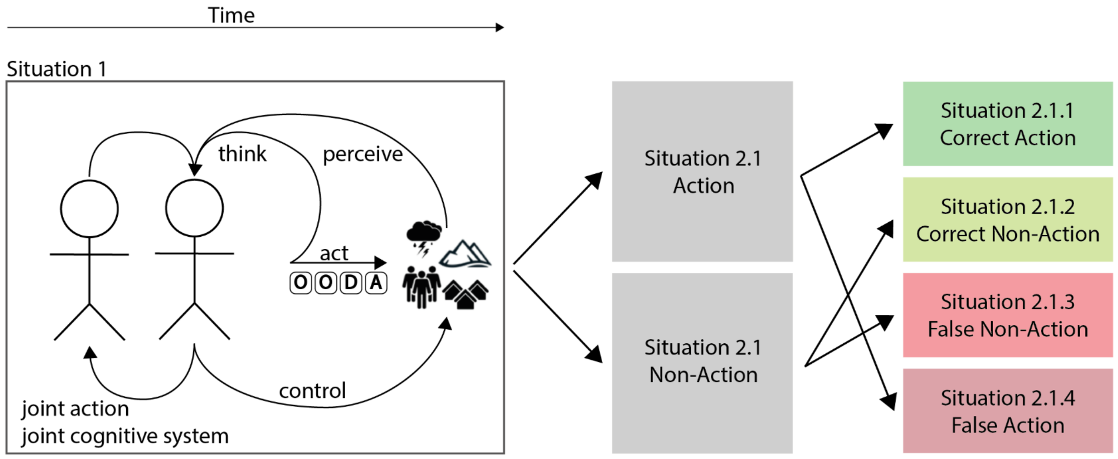


Figure 9: Joint thinking and acting as part of joint cognitive systems

Applied to real systems, nested control loops like the one in Figure 9, which allow joint action and joint cognition, play a role in almost every domain. In aviation, a crew works together to keep the airplane safe. A similar example are joint cognitive decisions between **fighter aircraft pilots and weapon system officers**, working at operational and tactical levels but with part of the activities overlapping (Pacaux-Lemoine & Loiselet, 2002). Another example is in Air Traffic Control, in which radar and planning controllers work together to detect and solve conflicts between aircrafts entering their controlled sector. Even if they work at different decisional levels (planning and tactical), some shared functions need joint thinking and acting (Pacaux-Lemoine et al., 1996).

Another **example** for joint action, joint cognition and cooperation of two or more humans is the **collaboration of a commander and a gunner** in a battle tank. A common situation is the hunter-killer mode, where a commander (the hunter) is using an independent sight to acquire and select the next target, which the gunner (the killer) afterwards shoots at. In this case, the commander has high situational awareness and works on the observe, orient and decide control loops, while the gunner is focused on the act loop for highest precision**.**

Step 4: Human-Machine / Human-AI system and teaming

While step 1 is inspired by Wiener (1954) and formulated in a very generic way that it could be any agent, animal, human or machine, who is perceiving, cognizing and acting, step 2 and 3 have been mainly described for humans so far. Step 4 is now focusing on teaming humans with machines, e.g. with an automation or an AI.

Theories about human-machine teaming date back to Wiener (1950). Milestones along this way were

- Licklider (1960) sketching a Man-Computer Symbiosis
- Rasmussen (1983) proposing the term cooperation
- Hollnagel and Woods (1983) and Sheridan (2002) sketching the design space
- Hoc and Lemoine (1998) and Hoc (2000) describing common ground and knowhow-to-cooperate as important parts of human-machine cooperation, exemplified for vehicle control by, e.g. Flemisch et al. (2003), Holzmann (2008), Flemisch et al. (2008), Hakuli et al. (2009) , Onken and Schulte (2010)
- Abbink (2006) working out a symbiotic driving, in cooperation with the haptic shared & cooperative control paradigm of H-Mode, e.g. Goodrich et al. (2006) or Altendorf et al. (2015).

As a gaze into the future of human-machine cooperation and symbiosis, Flemisch and Baltzer (2022) discuss reversibility and non-reversibility of human-technology/machine/AI symbiosis.

Regarding control, with increasingly complex human-machine cooperation it becomes more obvious that control and controllability are an essential aspect of successful human-machine cooperation, and that the sharing and trading of control is inducing a new complexity, but also a new degree of freedom compared to purely autonomous systems, controlled purely by the human, or purely by the machine. In parallel to the H-Mode and H-Metaphor concept, which describe a haptic coupling of human and machine, Griffiths and Gillespie (2004) generalized this into shared control as a concept where both human and machine are influencing the plant, coupled with a haptic device. Abbink et al. (2018) sketch a topology of such systems, Flemisch et al. (2019a) extend this design space to shared and cooperative control, where control can either be shared or traded, and where **control serves a higher purpose: to build up and maintain good cooperation**. For this challenge, Klein et al. (2004) describe 10 challenges for making automation a “team player”.

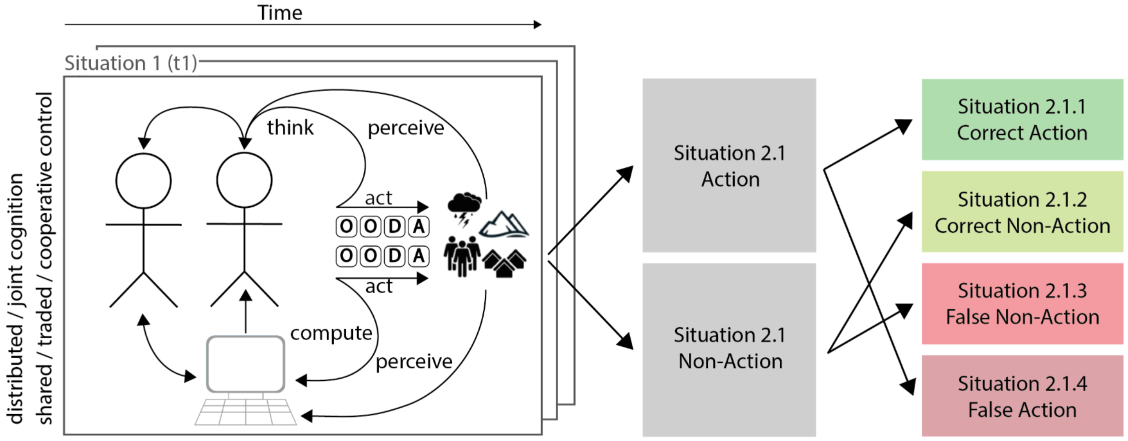

Time
Situation 1 (t1)
distributed / joint cognition
shared / traded / cooperative control
think
perceive
act
OODA
OODA
act
compute
perceive
Situation 2.1 Action
Situation 2.1 Non-Action
Situation 2.1.1 Correct Action
Situation 2.1.2 Correct Non-Action
Situation 2.1.3 False Non-Action
Situation 2.1.4 False Action

Figure: 10: Human-machine system and cooperation with other agents

Before we extend further, a look into the depth: Dual OODA Loop between human and machine

As we move from humans using software systems as tools towards humans **cooperating** as a team with intelligent systems, designers of such systems will have to consider how decision processes of the human and the machine will interact towards achieving the common goal that they share. A means of conceptualizing this is a dual OODA loop with the decision making processes of the human and the machine running in parallel. The nature of the functions that each agent performs and the interdependence between them will determine how important it is that these are synchronized. In this model we have used different terms within the two OODA loops to emphasize that the processes that humans and machines use are different, may occur at different speeds or are subject to different biases. Using such a conceptualization can be useful for considering how problems might arise in human-machine teams.

As the understanding how human and machine might cooperate becomes increasingly important, the model in Figure 11 also includes the Know-how-to-Cooperate on both sides, the human and the machine, and especially stresses the mutual connection via a common work space and interface, conceptualizing how coherence and mutually supportive behavior might be achieved.

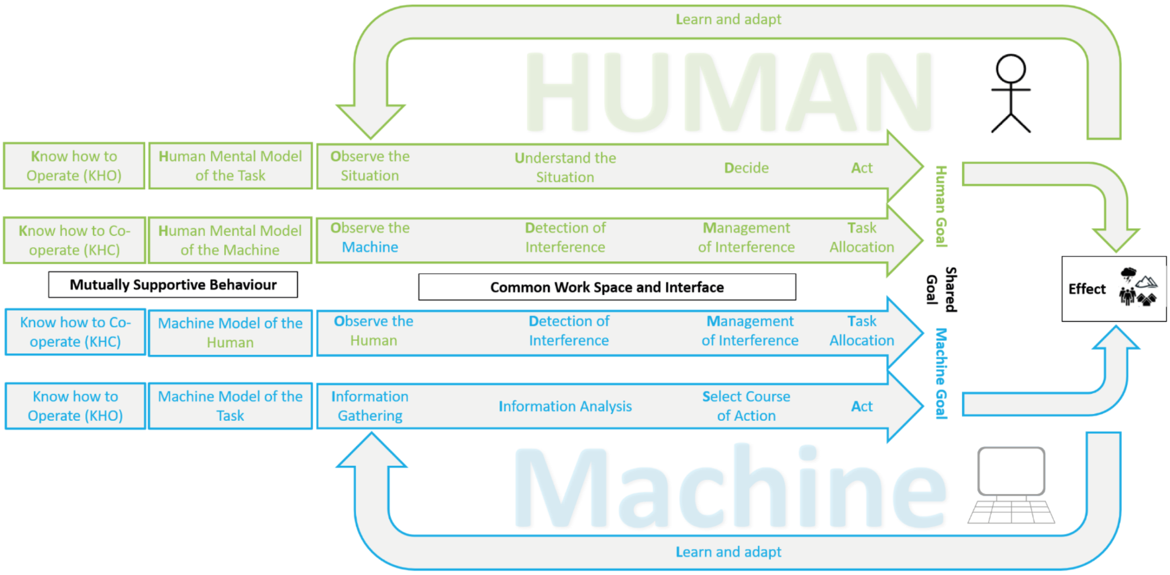

Learn and adapt
HUMAN
Know how to Operate (KHO)
Human Mental Model of the Task
Observe the Situation
Understand the Situation
Decide
Act
Human Goal
Know how to Co-operate (KHC)
Human Mental Model of the Machine
Observe the Machine
Detection of Interference
Management of Interference
Task Allocation
Mutually Supportive Behaviour
Common Work Space and Interface
Shared Goal
Effect
Know how to Co-operate (KHC)
Machine Model of the Human
Observe the Human
Detection of Interference
Management of Interference
Task Allocation
Machine Goal
Know how to Operate (KHO)
Machine Model of the Task
Information Gathering
Information Analysis
Select Course of Action
Act
Machine
Learn and adapt

Figure 11: More detailed model of human and machine agent cooperating with two "OODA" loops

Practically applied to the car domain, this could be the cooperation between a vehicle with high driving automation and the human driver. If drivers are still involved in the guidance tasks, e.g. maneuver guided driving, they need to interact with the machine, here an automation. In that case, communication between driver and automation must be enabled via a common work space, e.g. a bi-directional car display and a steering wheel where the automation senses the input of the human, and the human senses the input of the automation. This allows both a detection of interference, management of the interference e.g. by deciding about the task allocation, and the task allocation itself, e.g. by transitioning control between the driver and the automation. Transparency needs to be established e.g. of what the driver tries to achieve and how the automation acts upon that and a possibility to intervene on errors or oversights of the human or automation needs to be possible in a well working and cooperating team of driver and automated car.

Similar interaction and control patterns are applicable in the aviation or the production domain (e.g. Flemisch, Baltzer et al., 2022).

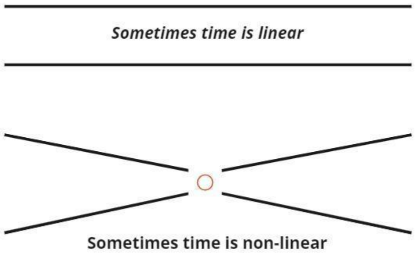
Step 5: Connecting a linear and a bow-tie understanding of **time**

Step 5 looks very innocent in the beginning, but might be one of the essential steps to connect the often tiny detail, which on the first glance causes our complex safety critical systems to be successful or to fail, to be successfully controlled or spin out of control, with all the important aspects that lead up to this success or failure, and might follow after success or failure. The mental operation necessary is one of the most challenging, at least for the part of our community who have been heavily trained in a linear, Newtonian understanding of time. On the other hand, it has been a part of the engineering community itself, who has developed this aspect of the model over decades, in lessons over lessons of success and painful failures.

In theory, at least in our Western view heavily influenced by Newton and other physicists, time is linearly expanding from the past stretching out into the future and is always similar. In more modern theories like the special theory of relativity, time is not linear, but influenced by changes in space (i.e. speed), and gravitation. For most people, this time model only applies to very small or quite big things, but not to normal life. In Greek philosophy and metaphysics, time not only in physics, but also in normal life has different qualities: Chronos, a more linear time, is often depicted with wings, because it flies fast, and with an hour glass and a scythe to chop off the mortals, when their time is over. Chairos however, the time of opportunity, is often depicted with winged shoes, flying bye fast, with a bald and slippery head, because he is difficult to grasp, but with a pigtail at his bald head so that we slow mortals have a chance to grasp Chairos at all.

We tend to smile at first about the wild imagination of our philosophical forefathers and mothers, but at a closer look we often discover why these stories have been passed on for millennia, and that the essence of the story might still be of precious value even in modern times. In regards to systems, many of us grow up educated only in linear systems, which applies only to very simple systems. With non-linear systems, it took a while to understand that a non-linear understanding of time could be valuable, and why certain points in time have more influence on the fate of the system than other times. **Bifurcation theory** describes how in these **bifurcations**, certain moments and situations in time, a small change to the parameter values can cause a sudden shift in its equilibrium, resulting in a big qualitative change in the systems behavior (e.g. Strogatz, 1995). **Examples** for this are the weather, where a butterfly flapping its wings could set off a tornado in Texas (Lorenz, 1963). A simpler version of this effect can be experienced with a chaos pendulum, where at certain points of the state space, small changes can make a big difference, nudging the system into a completely different behavior (e.g. Rubinsztejn, 2018).

For sociotechnical systems, it is so far not possible to describe these bifurcation points quantitatively, but especially in safety critical domains, they have been acknowledged and described qualitatively already for a long time. Starting from critical incident analysis in the military domain, the chemical industry developed a special view to connect linear time with bifurcations (e.g. Sneddon, 2017). This is achieved by plotting these events into a bow-tie diagram, which put hazardous events or critical incidents into the center of the analysis, and describe what leads up the event and what can be done to mitigate the consequences.



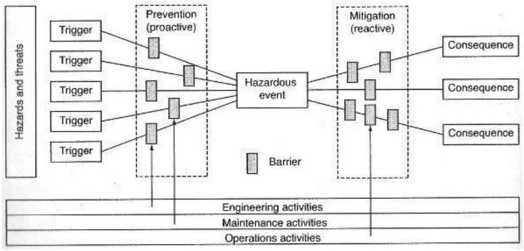


Figure 12a: Basic idea behind bifurcation theory
Figure 12 b: From Bifurcation Theory to Bow-tie diagram, example from Rausand (2013)

Applied to our challenge of a holistic model of control, we propose to use the bow-tie idea **not only** for analyzing accidents and incidents **in hindsight**, but also for understanding and designing our systems **in foresight** in a more general way. The idea is to take bifurcation theory as an inspiration, and to form our model into a bow-tie shape, placing **essential situations** as models of essential segments of reality into the center, and aspects leading up to these segments to the left, and aspects coming into play after the critical segments to the right (in Western writing styles. For other writing styles it could also make sense to let time flow from top to bottom). Segments can be situations e.g. described in use cases or vignettes, or specific subsystems.

Another look into the depth: Bow-tie and human-machine cooperation

Going further into the detail of a bow-tie relationship between humans and machines: As mentioned in steps 3 and 4, the enrichment of the model with human-human and human-machine cooperation, the next step of the model is the embedding and the description of the cooperative tasks. The types of functions described by the OODA loop have been also proposed by other models. Examples are a more detailed model proposed by the system engineering approach, i.e. the schematic map of the sequences of information processes (Rasmussen & Goodstein, 1987), or a similar model to the OODA model based on the four functions “Information acquisition”, “Information analysis”, “Decision selection” and “Action implementation”. This model was adapted to describe possible levels of automation, and so possible task sharing or trading between human and machine (Parasuraman et al., 2000). However, these models address individual functions of human and machine, when interacting with the situation to be controlled, but not the cooperative functions when they must gather information from the partner to do mutual adjustments to make common decisions and actions to reach their common goal. Those specific functions have been defined by Hoc and Lemoine (1998), they are part of the so-called Know-How-To-Cooperate. The last version of those functions are described in more detail in Pacaux-Lemoine (2020). For each individual function or group of individual functions of the OODA loop, cooperative functions aim at:

- Building up a representation (i.e. mental model) of cooperative partners (human or machine),
- Using this representation to try inferring their intentions and adapting own individual functions,
- Managing interferences or conflicts, when they occur, with the partners selecting the best solution and deciding function allocation,
- Using the Common Work Space to manage such cooperative objectives (Pacaux-Lemoine & Debernard, 2002).

The four cooperative functions, “Observation of others”, “Detection of interferences”, “Management of interferences” and “Function allocation” can be applied to each individual function of the OODA loop, or to some or all individual functions together. The more individual and cooperative functions are interrelated, the higher is the level of cooperation (Pacaux-Lemoine & Vanderhaegen, 2013). However, the higher the level of cooperation, the higher is the symbiosis between humans and machines and the associated risks like dependency or the loss of autonomy (Pacaux-Lemoine & Trentesaux, 2019).

Cooperation has its benefits if well managed. Too many negotiations may lower the advantages of cooperation if it is too difficult to know how to behave with partners. In addition, the more complex the human or machine organization is, the more difficult it is to find , and the more difficult it is to bring the decision making towards the “good” control of the situation. Figure 13 highlights such complexities with two main dimensions, “Complexity of human organization” and “Complexity of machine organization”. Several levels of complexity of cooperation are highlighted by this figure: cooperation between humans, cooperation between machines, cooperation within a system “one human, one machine”, and cooperation between several humans and several machines. Moreover, humans and machines may not be at the same decisional level, and so not at the same layer of cooperation, as further described in Pacaux-Lemoine and Flemisch (2019).

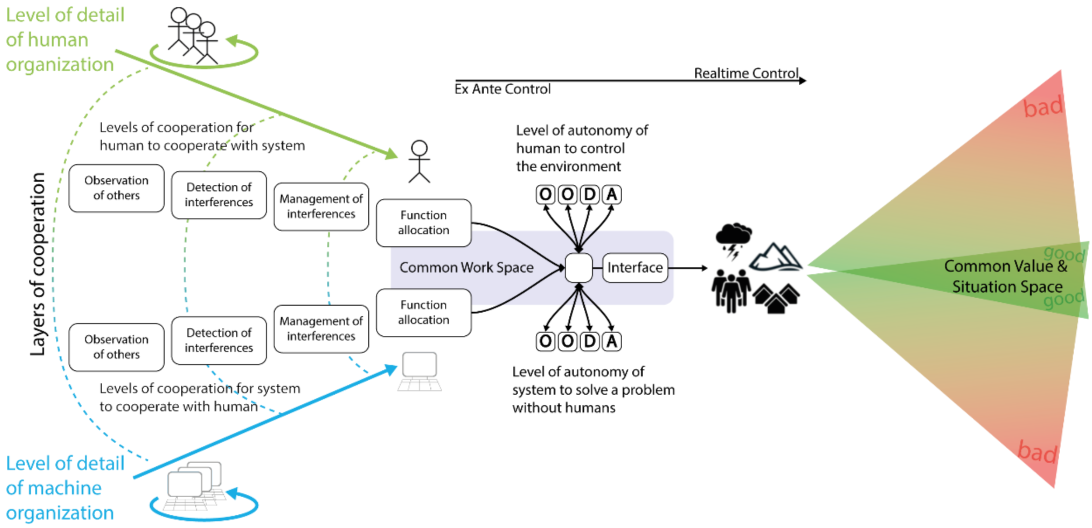

Level of detail of human organization
Realtime Control
Ex Ante Control
Levels of cooperation for human to cooperate with system
Level of autonomy of human to control the environment
Layers of cooperation
Observation of others
Detection of interferences
Management of interferences
Function allocation
O O D A
bad
Common Work Space
Interface
good
Common Value & Situation Space
good
Observation of others
Detection of interferences
Management of interferences
Function allocation
O O D A
Levels of cooperation for system to cooperate with human
Level of autonomy of system to solve a problem without humans
bad
Level of detail of machine organization

Figure 13: First steps towards a Bow-tie model of human-machine cooperation, adapted from Pacaux-Lemoine and Flemisch (2019).

Figure 13 proposes several possible positions for the “good” vs. “bad” situation (represented by two triangles). The objective is to explain what could be the impact of a decision made only by humans (upper triangle), or only by machines (lower triangle) or from a successful cooperation (overlapped area). The overlapped area, the so-called “Common Value & Situation Space”, is an extension of the “Common Frame of Reference” proposed by Hoc and Lemoine (1998). This space makes the evaluation of situations possible according to predefined values.

This relationship between humans and machines is not per se symmetric: As machines are not limited by biological limits e.g. by biological nerves, it is very likely that they become much faster. This difference in timing may bring challenges to good decision making because the human might be tempted to follow the machine. Moreover, meaningful control may also be different from human to machine if we consider that machine could be myopic, less opportunistic and less adaptable to unexpected events. Figure 13 shows how to mix the capabilities of each side (human and machine) by the means of cooperation. For this human-agent activity, Klien et al. (2004) already proposed ten challenges to make the automation a real team player, of which only a couple are solved.

In order to address those challenges and the relative complexity of the humans-machine organization, human-machine system design and evaluation methodologies exist and support designers, engineers and stakeholders to reach their common objectives. Examples of principles have been proposed by Pacaux-Lemoine and Flemisch (2022) like the iterative procedure with spinal refinement, the Human-Cyber Physical System Integration method and the experimental dimensions highlighting the necessary correlation between the successive improvements of the experimental conditions, systems' capabilities and human training.

Step 6: Connecting the detail with the whole

While the bow-tie shape already allows to bring the most critical aspects into the center of the model, and describes what aspects lead up to important situations, and what happens after these situations, Step 6, may be even more challenging for some minds, but is the most crucial. It goes a step further: The goal is to connect the small with the big, the detail with the whole, without losing the utility of the model.
In order to understand why this could be challenging for some readers, it might help to understand how science developed over the last centuries in an interplay of understanding the detail and understanding the whole. The attempt to understand everything together is still visible in Greek philosophy (Burke, 2015), and in the original meaning of the name university, which comes from Latin: "universum" = "all things, everybody, all people, the whole world", + "versus"(past participle of vertere) = "to turn, turn back, be turned, convert, transform, be changed." (Online Etymology Dictionary, 2022). As Peter Burke described, with increasing specialization necessary to understand more details of this world, and not only to understand, but also to influence (as in liberal and other arts or in engineering, which was considered an art until the late 19th century) and to build part of this world (as in engineering), a second culture emerged which values precision in the detail, but also has difficulties with a more holistic perspective (Burke, 2015). Snow and Collini (1956)described this as "two cultures" and argued that only in a fruitful connection of these two cultures humankind can proceed. This tension field has been discussed many times e.g. in the *Methodenstreit* of 1890 at German universities, but also in modern times, e.g. regarding the challenge to balance security and freedom in cyberspace, which are a "culmination of absence of dialogue between the two cultures" (Ilves, 2014). In the beginning of the 21st century, it becomes increasingly clear and accepted that a **proper combination of cultures**, with sufficient specialization on the one hand, and an interdisciplinary and multidisciplinary dialogue on the other hand, might be the most fruitful way to go. With this motive in mind, our model expands from the world of effective human control and human-machine systems, to the bigger perspectives of system of systems, organization, societies and the Earth as a whole. With that, similar to a gym class, where we learn to stretch our body, this model might also motivate the stretching of our minds, step by step.

System theory has always strived to provide a theory, which can bridge the tension field between smaller details and the whole. Philosophy already structured this tension field e.g. from Aristotle ("The whole is greater than the sum of its parts" (Ross, 1924)) up to Esfeld (2001), who describes atomism versus holism in social sciences. For an interesting discussion on the Chinese concept of Tianxia 天下 "(all) under heaven" see Zhao Tingyang (2010), for an application to cognitive warfare refer to Flemisch (2022).

More technically, Haberfellner et al. (2019), who successfully applied system theory to systems engineering, describe analysis and synthesis as two different but complementary thinking directions. This combination of analysis and synthesis enables to dissect complex systems into a lower level of detail into subsystems and their relationships, and to synthesize them back into bigger systems, and systems of systems. In environmental sciences, there is a strong movement towards considering the whole environment (e.g. with an Elsevier Journal “Science of the Total Environment”), which strives for holistic models. An example for this is the research of (Baratsas et al., 2021; Stefanos et al., 2021) on foodchains, who demand a holistic systems engineering approach.

We are aware that after centuries of developing a science of the details, the science of the whole might not be developed to a similar extent yet, and there is still the necessity to develop theory even further. On the other hand, the call for integrating smaller with bigger perspectives gets increasingly heard and answered in the communities (e.g. Elgezabal & Schumann, 2012; Jones et al., 2005), and is also visible in the huge movement towards a Human Systems Integration (e.g. Flemisch et al., 2021; Shea, 2019). There is already substantial theoretic ground to stand on. More importantly, with control and controllability, there are connections between small human-machine systems, and bigger systems of systems and societies that are so crucial, that without considering these connections, no model would be able to make a real impact on the real world.

Let’s look at the word holistic a bit closer, and derive a practical definition of what we mean with a holistic model: “holistic” means “dealing with or treating the whole of something or someone and not just a part” (Cambridge Dictionary, 2022a). How big could “the whole” be? It becomes immediately clear that “the world [as] everything, that is the case”, the famous opening words of Wittgenstein’s Tractatus Logico-Philosophicus, describes the upper limit, but might not look practical on the first glance. We might be tempted to define a border, a cut off line of level of detail or wholeness, just for practical reasons. Instead of total holism, we might be tempted to work on (at least for modeling, system science, systems engineering and human systems integration) a more commensurate holism:
A perspective, which considers the whole of a system and its relationship to its environment up to a scale beyond which no substantial effects are expected by the analysts.

Applied to our challenge of control and Meaningful Human Control, we see an increasing trend towards global systems, e.g. in defense and economic systems, with phenomena like flash crashes and flash wars, that it would not make sense to have a holistic model smaller than the globe of our planet Earth. Now comes an important fork in the development of our thinking: Looking at bigger scales like the solar system, our galaxy, our galaxy clusters etc., we might not see connections yet with these bigger systems, so we might be tempted to choose the cut-off line with planet Earth. On the one hand this would keep the freedom to extend the commensurate horizon anytime if necessary, e.g. if we need to consider the energy production of the sun or additional resources at other planets of our solar system. However, we would advise not to limit ourselves to system earth: Also Earth is embedded in a larger system of systems, and e.g. heavily influenced by the sun. Moreover, Homo Sapiens has always been a species, which is exploring and expanding to new horizons. It makes a difference, whether we consider ourselves as a species bound to planet Earth, or a species, which is able to look, and since many decades to even travel to space beyond Earth. In order to be truly holistic, we should include this horizon beyond Earth at least with one extra box “space”, and connect this with the immediate perspectives of our local systems: “Keep your eyes on the stars and your feet on the ground” (Theodore Roosevelt, 1900).

Between a global perspective of our global environment Earth embedded in space, and a local perspective of a human-machine system, what intermediate layers should we choose? Going beyond the “classical” human-machine system(s), it becomes increasingly fruitful to define systems of systems, where individual systems are being joined and “deliver important emergent properties, which have an evolving nature that stakeholders must recognize, analyze and understand”. Maier (1996) e.g. describes five traits of systems of systems:

- Operational independence of elements;
- Managerial independence of elements;
- Evolutionary development;
- Emergent behavior;
- Geographical distribution of elements.

System of systems approaches started in the late 1990s, and are now sufficiently matured to be quite useful for understanding and designing larger scale combinations of systems.

A big progress of systems engineering was the integration of organizational perspectives, forming Human Systems Integration. It looks like beyond the system of systems, a fruitful layer could be organizations. Organizations have been subject to research already for decades. Weick (1974), as one of many examples on organizational sciences, e.g. describes organizations as follows:

> "The word, organization, is a noun and it is also a myth. If one looks for an organization one will not find it. What will be found is that there are events, linked together, that transpire within concrete walls and these sequences, their pathways, their timing, are the forms we erroneously make into substances when we talk about an organization (Weick, 1974, p. 358).

Weick’s work is also interesting because he and his colleagues formulated a concept of **sensemaking** through organizations, here as a process that is grounded in identity construction, retrospective, cue extraction, dialogues and story-telling and plausibility checking. Weick (1993) explicitly suggests that the focus of organizations should shift from decision making to how to elicit **shared meaning,** more as a principle-based than a rule-based approach (Helms Mills et al., 2010).

It becomes increasingly clear that the “Meaningful” in MHC could be individual meaning, but even more an organizational shared meaning. Further, it also considers issues beyond the organizational level, more on a societal level, with all its social, cultural, legal, ethical and moral aspects. **Societies** are usually considered as a large group of people who live together in an organized way, making decisions about how to do things and sharing the work that needs to be done (Cambridge Dictionary, 2022b). Rahwan (2018) argues that it is necessary to have the “society-in-the-loop” for any algorithms that have a broader implication that surpasses individual humans or organizations. In other words, it is not enough to have a given human “in the loop” or in control. This shift raises multiple questions such as how to balance the competing interests, values, and norms of different people, organizations, and governments (Gabriel, 2020).

As many of our complex systems do not only affect a single society, but multiple societies and their interconnectivity, no model of such a system would be sufficiently capable without a larger layer: Our **planet Earth**, which, at least at the beginning of the 21st century, hosts all societies and states. “Global” is the adjective which is widely used, and it includes environmental, political, economic and other implications.

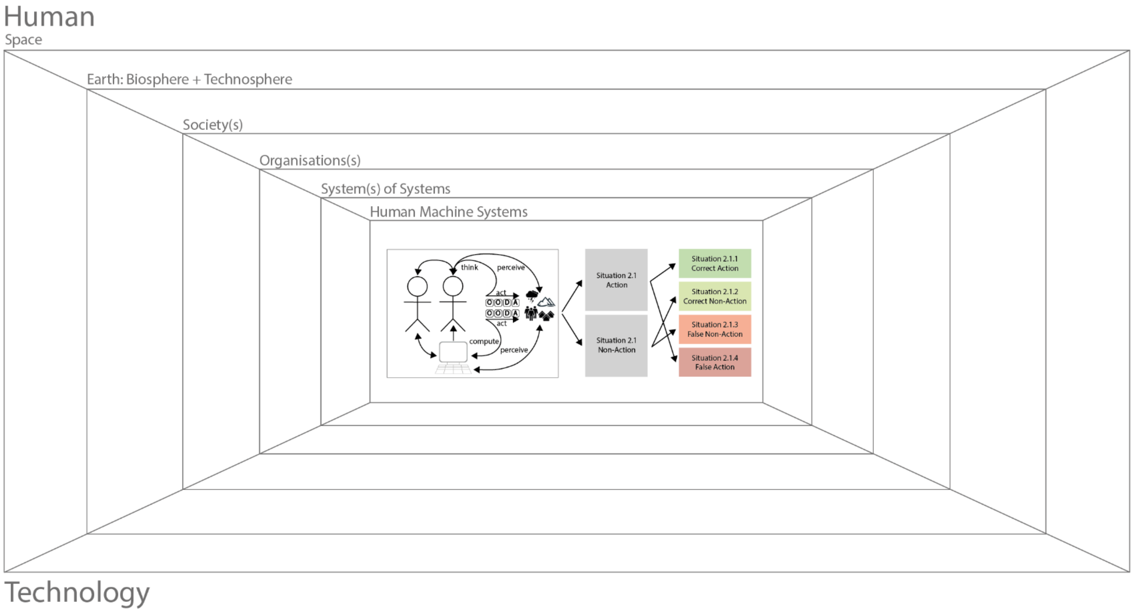


Figure 14: Holistic bow-tie model for control in human-machine / human-AI systems

Overall, these nested layers from the global perspective to a local human-machine system form a holistic bow-tie diagram, placing the most critical situations in the middle, and nesting them with the most influential layers of meta-systems, e.g. system-of-systems, organizations or societies (Figure 14), so that chains of interrelationships can be shown also across different levels of details (see Figure 14).

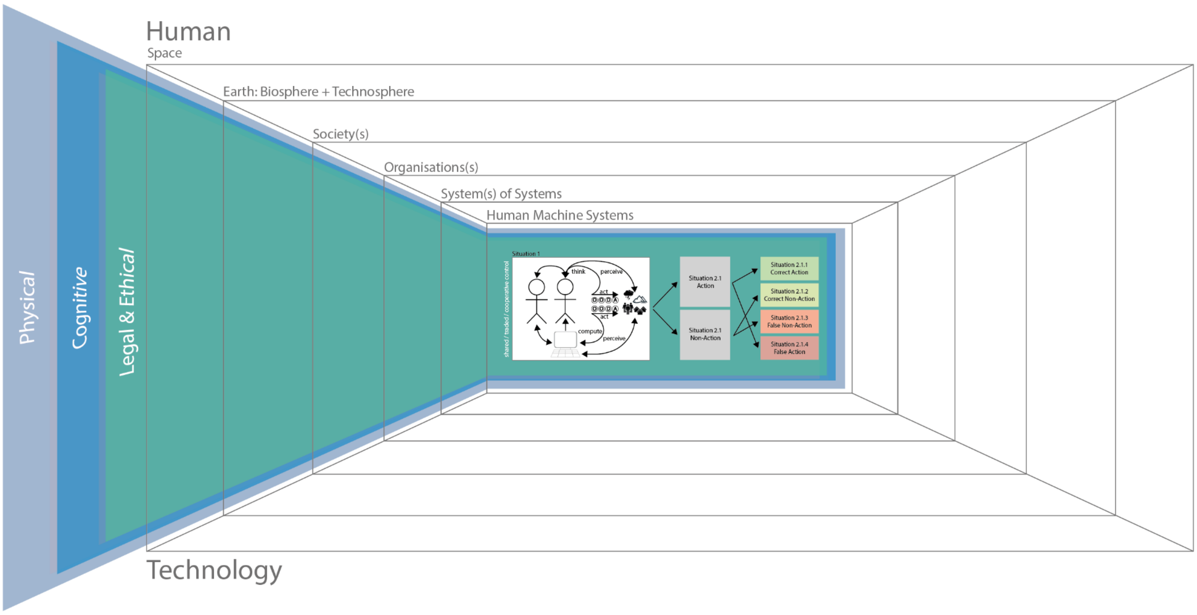


Figure 15: Holistic bow-tie model for human-machine / human-AI systems with transversal layers, here of physical, cognitive, ethical and legal dimensions

Figure 15 shows an extension of the holistic bow-tie model, where transversal layers are connecting the layers of detail. We propose a layered approach, starting with a physical layer, which describes the physical interrelationship of these layered systems, and based on this a cognitive layer, which describes all cognitive functions and interrelationship of these layered systems. This looks trivial at a first glance, but already contains major insights for which science has struggled for centuries:

- Physical and cognitive layers are not separated subsystem, but the cognitive layer is based on a physical layer. This reflects the discussion of e.g. Damasio (1994), who describes how human cognition is based on a physical and physiological layer, in which e.g. the brain thinks and judges using responses of the body (somatic markers) (Damasio, 1994). The layered approach also allows to describe bi-directional feedback loops between cognitive and physical aspects.
- Cognition is not limited to a singular human brain. Computers can be part of cognitive processes as well, and form, together with humans, organizations and societies, joint cognitive systems, as already described earlier. Moreover, Cognition can also be thought as distributed cognition along all layers, as described by Hutchins (1995).

At the beginning of the 21st century, there is a strong movement not only towards including the physical and cognitive layer, but also an ethical layer, e.g. as value based engineering (e.g. Spiekermann & Winkler, 2020) or as ethical systems engineering (e.g. Gillespie, 2019). It is important to note that to be truly holistic, other fundamental aspects of human existence like a spiritual layer might also be included, but are, in the 2020s, outside of the paradigm of science consciously chosen for this publication.

One more look at the aspect of time: direct and indirect control, operational, tactical and strategic control

Looking at MHC with a holistic bow-tie model-lens, it is crucial to connect not only layers from the smaller to the bigger systems and vice versa, but also to connect our different understanding of time. The bow-tie part of the model puts the individual human and human-machine system into the center, embedded and constantly evolving from the past, constantly developing into the future. With the lens of a physical understanding, the present might be an infinitely small part of a continuum. Looking with the lens of human systems integration, the present is more what we as Homo Sapiens **perceive** as present, as a connected moment, in a certain situation.

From the original meaning of control and controllability, e.g. as Wiener (1950) and his colleagues understood it, the main part of control is by feedback loops from the actual present, which becomes the past, into a future present, which becomes the actual present. The "back" in "feedback" is meant as "from the effect of an action back to the perception and action", rather than back in time. Only with learnable systems, the past starts to influence the present even more in the form of a feedforward system. Looking at control and MHC, it becomes increasingly clear that we might have to differentiate several ways to influence and control:

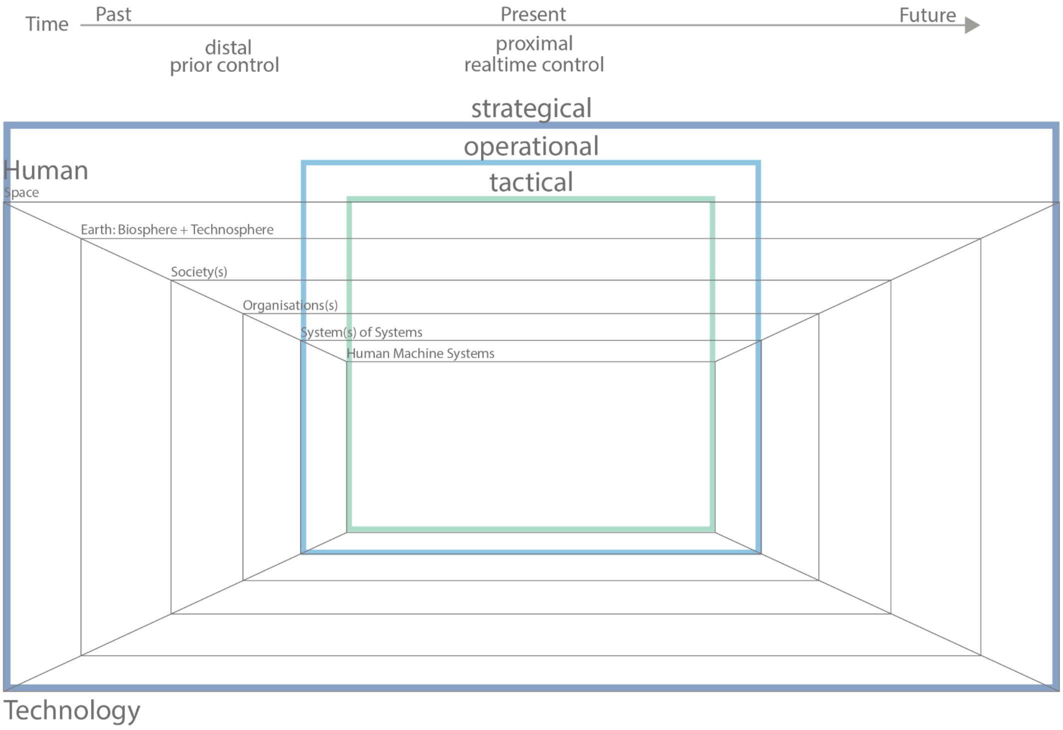


Fig 16: Direct and indirect control; tactical, operational and strategic control

Figure 10 - 15 show control in presence or **direct control,** where the control loops are closed in the presence. An **example** on the level of the human-machine system would be a driver directly controlling the vehicle in the here and now, ready to react anytime. An **example** of direct control over several layers would be a war room like in the “Situation Room” in the US White House being directly connected to a war site, e.g. the location of Osama Bin Laden, giving a Commander in Chief the opportunity to abort a mission in the present moment.

Concerning control from the past, i.e., prior control, or **indirect control**, the control loops are mainly feedforward loops, which influence actors or agents in a future presence. An **example** are rules of engagement, which are formulated at a certain time often by an interplay of societies and organizations, then being trained to individuals or embedded into technical systems like computers or AI, then being used in certain situations, and then, after-the-fact being evaluated e.g. at court.

What looks trivial at the first glance can become quite complex at the following glances: There were and still are different interpretations of whether the past completely determines the future, or whether there is something like chance or freedom. While one of our brightest brains believed in "Gott würfelt nicht" ("God does not play dice", Einstein 1926 in Baggott (2018)) modern system and chaos theory dared to think beyond Einstein. It becomes increasingly clear that in complex systems of systems, emergent effects exist, which cannot be forecasted by the characteristics of the individual systems. Even without emergent effects, in a perfectly mechanistically describable system like a chaos pendulum, deterministic chaos is a fact, where already after a short period of time, it becomes practically impossible to forecast the outcome of the system. It becomes increasingly clear that with increasing time distance and complexity of the systems and situations, indirect control is certainly becoming non-deterministic to a point, that a majority of stakeholders, independent in what God, Gods or concepts they believe in, would no longer speak of neither effective nor meaningful control.

One way out of this dilemma is to limit the application of indirect control to certain geographic areas, and describe this as an **informed, conscious decision**, e.g. by Horowitz and Scharre (2015) e.g. to ensure the lawfulness of an action. Despite the fact that it is always good to make a conscious and informed decision, this argument could run into a deadlock if a law would require not only effective control, but also Meaningful Human Control.

What might be **a way out of the dilemma** of indirect control is to differentiate according to the different critical time constants in systems, as was done for centuries with the terms **strategic, operational and tactical** (see Figure 16). These terms do not only describe a different geographic scope, but also a different scope in time: Tactical is the most direct control, operational control still has a focused area and time frame, but is usually not connected in real time feedback loops, while strategic has a much wider and longer scope. Flemisch et al. (2017) describe such a layered model which connects the strategic, tactical and operational perspective with the concepts of cooperative, shared and traded control, with a transversal perspective on cooperation, therefore allowing to effectively “joining the blunt with the pointy end of the spear" or forging a complex chain from society to organizations to system-of-systems, human-machine systems and individuals.

It is essential to keep in mind that the holistic bow-tie diagram is not a time precise diagram like some physical diagram, but a rough mindmap with the main goal to show causal relationships between humans and technology and the different layers of details. This leads up to the essential situations, which determine the fate of the human-machine / human-AI system and their nesting systems of systems, organizations and societies. An application example for this is shown further down.

Holistic bow-tie model in action: selected interrelationships

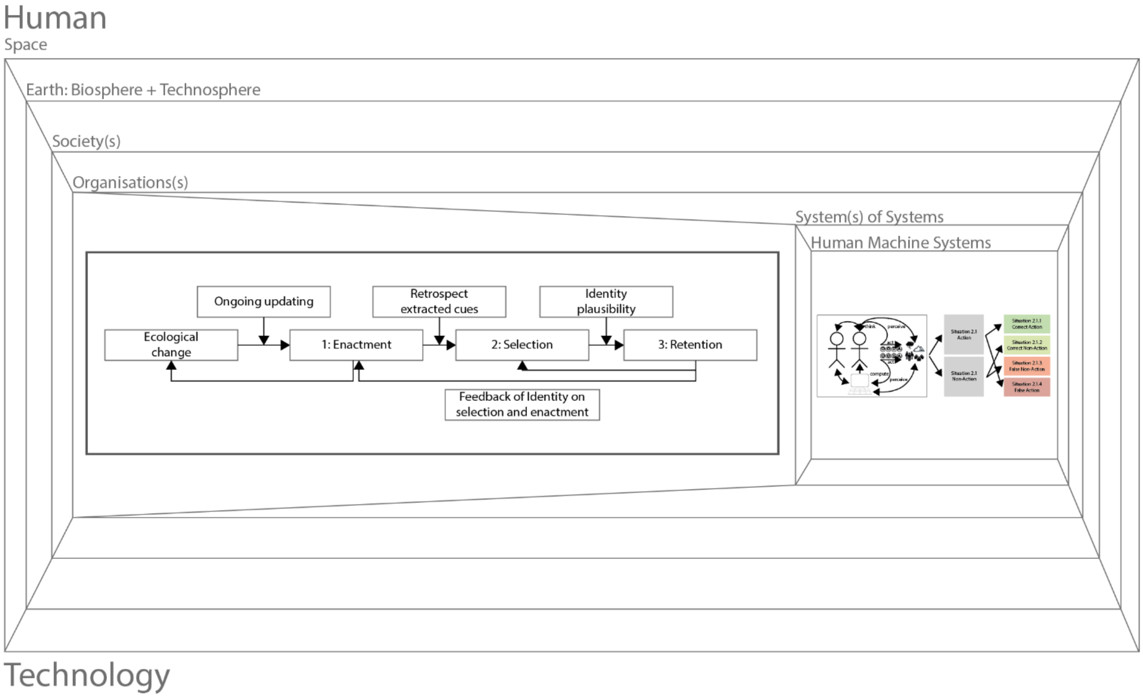


Figure 17: Different configurations of the same holistic bow-tie model: adapt focus, but never skip holistic relationships completely. (Example for an organizational issue on sensemaking (Stieglitz et al. (2018) adapted from Weick et al. (2005))

Figure 17 shows an example of how we envision this holistic bow-tie model to be used. If the focus needs to be placed on a different layer, the diagram can be adapted, flexed and stretched, so that there is always enough space to show the aspects in focus. In contrast to non-holistic system diagrams, the key is not to cut away the other layers, but always keep them to motivate the mind to also think holistically, i.e. especially consider the connections from the bigger to the smaller system perspectives.

Figure 18 shows an alternative use of the holistic bow-tie diagram, where the detail is "zoomed out" from the overview diagram in a way that the connection of the detail and the holistic perspective is easily maintained.

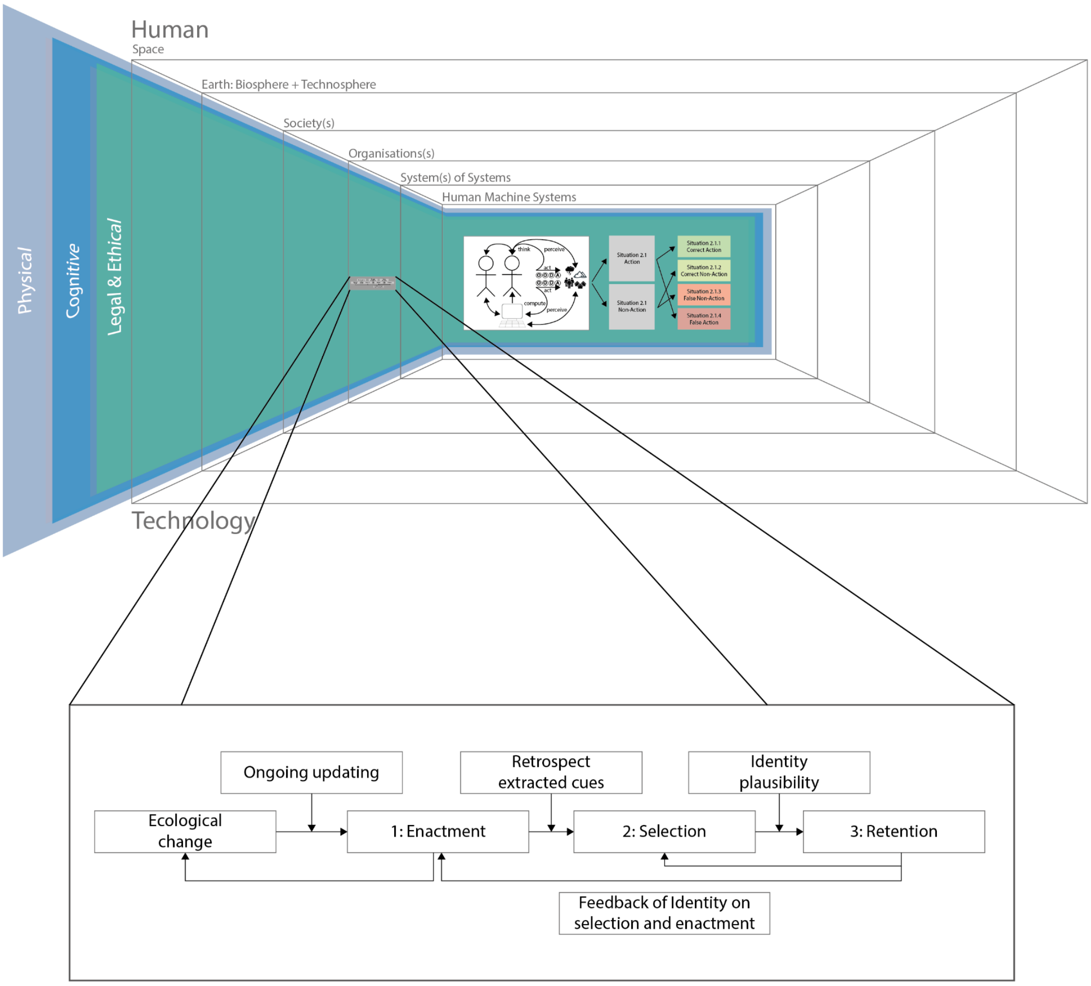


Figure 18: Alternative 2 of keeping the connection between the big and the small by zooming out (Example for an organizational issue on sensemaking, Stieglitz et al. (2018) adapted from Weick et al. (2005))

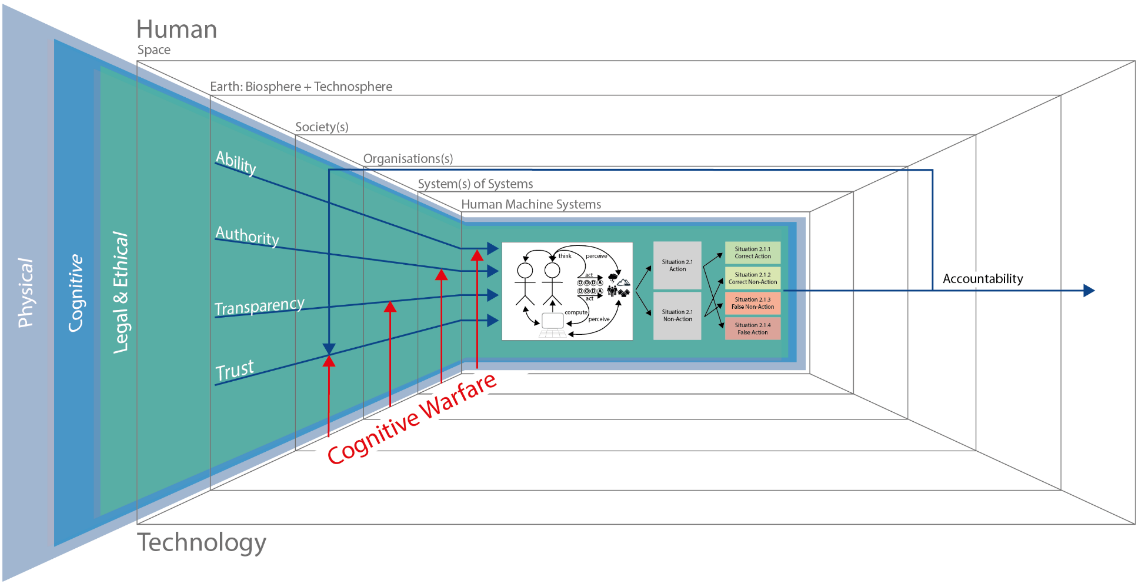


Figure 19: Example for the holistic bow-tie diagram applied to cognitive warfare (Flemisch 2022)

Figure 19 shows an example how the holistic bow-tie diagram can be used not only for MHC, but also for other issues. Here the example of cognitive warfare or cognitive defense is presented, where an adversary is attacking (red arrows) the transversal cognitive layer and its processes (blue arrows), here of NATO, by corrupting trust, degrading transparency, undermining authority and thereby degrading cognitive abilities of joint action e.g. in combat or cyber defense. More details can be found in Flemisch (2022).

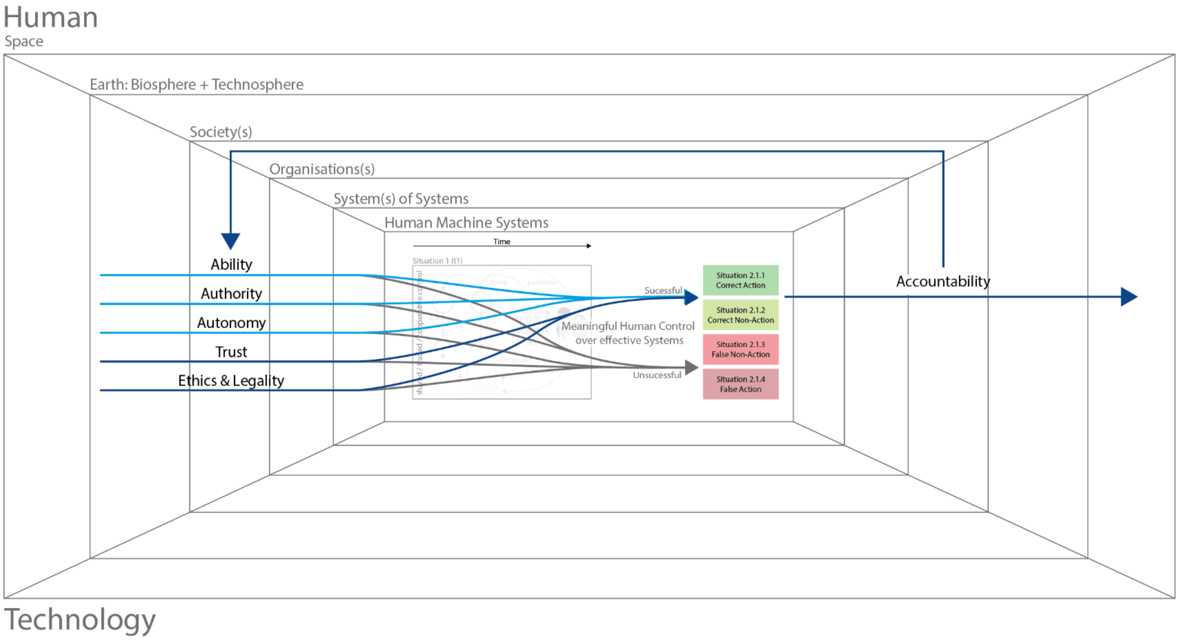

Human
Space
Earth: Biosphere + Technosphere
Society(s)
Organisations(s)
System(s) of Systems
Human Machine Systems
Time
Situation 1 (t1)
shared / traded / cooperative control
Ability
Authority
Autonomy
Trust
Ethics & Legality
Sucessful
Unsucessful
Meaningful Human Control over effective Systems
Situation 2.1.1 Correct Action
Situation 2.1.2 Correct Non-Action
Situation 2.1.3 False Non-Action
Situation 2.1.4 False Action
Accountability
Technology

Figure 20: Holistic bow-tie diagram of Meaningful Human Control over effective systems, as chain of ability, authority, autonomy, control/controllability and accountability

Back to our challenge of effective and Meaningful Human Control: Figure 20 shows an example how the holistic bow-tie diagram can be applied to our challenge of effective and Meaningful Human Control. Effectiveness based on sufficient ability is an important prerequisite of control, and necessary so that control can be successfully applied. Meaningful Human Control goes much further, and asserts that this control also has a positive meaning or sense for the involved stakeholders in the different system layers, from the local agent all the way through the system of systems, organizations, societies and the global environment. It is important to note that this is a relative concept, not an absolute: Santoni de Sio and van den Hoven (2018) argue that a system can be under MHC with respect to some humans, but at the same time result in behavior unacceptable by other humans. Whenever we talk about MHC, we also have to talk about the involved stakeholders.

Meaningful Human Control also involves a meaningful interplay of ability, authority, autonomy, control and finally accountability, respecting the double and triple binds between these concepts (Figure 21). Examples for this are the unsafe valley of AI and automation (Flemisch et al., 2017) or the moral crumple zone (Elish, 2019), where humans are made accountable, but do not have sufficient abilities to really control the artifact.

Both effective and Meaningful Human Control, or better Meaningful Human Control over effective systems, also need good situation awareness, a not-too-high and not-too-low workload, and calibrated trust, in order to develop enough ability, which enables control. MHC also needs a minimum of autonomy; otherwise it would not make sense to speak about control. Responsibility, e.g. the feeling of one of the agents to be responsible, motivates this agent to strive more MHC.

Only people who have enough ability, autonomy and authority to control a certain situation should be called to have Meaningful Human Control, for which they can be made accountable. More details about these brittle interrelationships can be found in Cavalcante Siebert et al. (2022), building on earlier work by Flemisch et al. (2012).

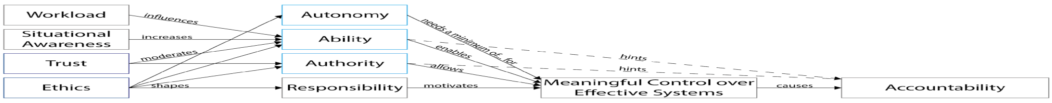


Figure 21: Double and triple binds of authority, ability, autonomy, responsibility, control and accountability, applied to MHC (extension from Flemisch et al. (2012))

Outlook: Where could we go from here?

With Figure 20 and Figure 21, it becomes increasingly clear that MHC is more specific than control, and more than what traditional control theory describes as efficient system control. MHC puts more demands on the system design than just looking for efficiency. Meaningful means more than effectiveness. Nevertheless, effectiveness of control, and effectiveness of the systems remain an important aspect in all life cycles of these systems. We propose to look at effectiveness of systems and Meaningful Human Control as two sides of one coin which belong together. We would like to bring this together with the term “ MHC over effective systems”. While the technical community and the human factors community have been mainly focused on the effectiveness, the ethical, legal and increasingly also the political community have an increasing emphasis on the sensemaking (Weick, 1995) or meaningfulness of control. Only if these parts of the community work cooperatively together, we will achieve Meaningful Human Control over effective systems.

As MHC is even more widespread over the different layers of detail in the holistic system of systems, and taking Wittgenstein serious in his criticism of classical definition theory being too rigid for complex concepts (e.g. Swartz, 1997), it makes more sense to define MHC as a family concept or cluster concept, where a number of relationships between the (effective) control and other factors, e.g. of ethical alignment, enough autonomy etc. are well balanced. While a couple of these relationships are already sketched e.g. in this first issue of the Edward Elgar Research Handbook, many research questions especially on the complex interplay of these system qualities, and especially on the interplay of the stakeholders, and of the interplay of MHC with other essential concepts and qualities are still open to be investigated.

The goal of the model is to enable a better balance between the different perspectives, and especially enable to link these perspectives together. The proof of the model is in its use, which could also be cooperatively: We share the source files of the figures on a Microsoft SharePoint (HolisticModelDatabase, 2022) free to use, to modify and to redistribute (with credits, using a creative commons Attribution –ShareAlike CC BY-SA). If you want to share your models in return on the same database, just contact us.
Scientifically there is a lot to discover and to gain, mapping these interrelationships into a holistic perspective, putting more and more light on the delicate interplay of humans, technology and organizations, societies and our global environment from the tiny detail, which can make the difference between failure and success, to the big picture, in which we all live.

From the first spears to our complex systems of today, we have come far, always by developing and expanding our physical, cognitive and ethical abilities. There is no doubt that there are global problems, which could easily ‘evolve’ Homo Sapiens into Homo Extinctus, taking many other species into extinction as well If we want to influence or even control our own fate, it could be helpful to understand “sapiens” less as a fact, but more as a potential, chance and obligation. As we already have tremendous physical abilities that have serious effects on a global scale, let us not hesitate to also expand our cognitive and ethical abilities up to a global scale where it can cope with the global challenges. Time is up.

Acknowledgements

The models here were shaped in fruitful discussions not only of the authors, but with a couple of other groups and entities, e.g. the NATO-STO Human Factors & Medicine panel, and especially with the colleagues of NATO STO RTG HFM-330 “Meaningful Human Control over AI-based systems”. Special thank you to them, and also to Gen. Gäbelein and Col. v. Rauchhaupt, for their invaluable insight into the “chain of trust”, on which we all in NATO and on this planet depend. Special thank you for Michael Preutenborbeck for the invaluable support and extra hours of editing.

References

Abbink, D. A., Carlson, T., Mulder, M., Winter, J. C. F. de, Aminravan, F., Gibo, T. L., & Boer, E. R. (2018). A Topology of Shared Control Systems—Finding Common Ground in Diversity. *IEEE Transactions on Human-Machine Systems*, *48*(5), 509–525. https://doi.org/10.1109/THMS.2018.2791570

Abbink, D. A. (2006). *Neuromuscular analysis of haptic gas pedal feedback during car following*. Delft University of Technology.

Altendorf, E., Baltzer, M., Heesen, M., Kienle, M., Weißgerber, T., & Flemisch, F. (2016). H-Mode: A Haptic-Multimodal Interaction Concept for Cooperative Guidance and Control of Partially and Highly Automated Vehicles. In H. Winner, S. Hakuli, F. Lotz, & C. Singer (Eds.), *Springer* *reference. Handbook of driver assistance systems: Basic information, components and systems for active safety and comfort* (pp. 1499–1518). Springer International Publishing. https://doi.org/10.1007/978-3-319-12352-3_60

Altendorf, E., Baltzer, M., Kienle, M., Meier, S., Weißgerber, T., Heesen, M., & Flemisch, F. (2015). H-Mode 2D: Eine haptisch-multimodale Bedienweise für die kooperative Führung teil- und hochautomatisierter Fahrzeuge. In H. Winner, S. Hakuli, F. Lotz, & C. Singer (Eds.), *Handbuch Fahrerassistenzsysteme* (pp. 1123–1138). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-05734-3_60

Baggott, J. (2018). *Quantum space: Loop quantum gravity and the search for the structure of space, time, and the universe*. Oxford University Press.

Baratsas, S. G., Pistikopoulos, E. N., & Avraamidou, S. (2021). Circular Economy Systems Engineering: A case study on the Coffee Supply Chain. In M. Türkay & R. Gani (Eds.), *Computer Aided Chemical Engineering. 31st European Symposium on Computer Aided Process Engineering* (Vol. 50, pp. 1541–1546). Elsevier. https://doi.org/10.1016/B978-0-323-88506-5.50238-2

Billings, C. E. (1996). *Aviation Automation: The Search for A Human-centered Approach* (First edition). *Human factors in transportation*. CRC Press.

Boardman, M., & Butcher, F. (2019). *An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It*. Porton Down. https://www.sto.nato.int/publications/pages/results.aspx?k=STO-MP-IST-178%207%20-%201%20An%20Exploration%20of%20Maintaining%20Human%20Control%20in%20AI%20Enabled%20Systems%20and%20the%20Challenges%20of%20Achieving%20It&s=Search%20All%20STO%20Reports

Boillot, L. (2014). *Key areas for debate on autonomous weapons systems: Memorandum for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*. Geneva. https://article36.org/updates/key-issues-for-debate-on-autonomous-weapons-systems/

Boyd, J. R. (1996). *The essence of winning and losing*. Bluffton, South Carolina. https://ooda.de/media/john_boyd_-_the_essence_of_winning_and_losing.pdf

Burke, P. (2015). *What is the History of Knowledge?* John Wiley & Sons.

Cambridge Dictionary. (2014). *control*.

Cambridge Dictionary. (2022a, August 16). *holistic*. https://dictionary.cambridge.org/dictionary/english/holistic

Cambridge Dictionary. (2022b, August 16). *society*. https://dictionary.cambridge.org/dictionary/english/society

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2022). Meaningful human control: actionable properties for AI system development. *AI and Ethics.* Advance online publication. https://doi.org/10.1007/s43681-022-00167-3

Chamberlin, J. E. (2006). *Horse: How the horse has shaped civilizations*. Bluebridge.

Chamberlin, J. E. (2010). *Horse: How the horse has shaped civilizations*. Vintage Canada.

Chen, J. Y. C., Flemisch, F. O., Lyons, J. B., & Neerincx, M. A. (2020). Guest Editorial: Agent and System Transparency. *IEEE Transactions on Human-Machine Systems*, *50*(3), 189–193. https://doi.org/10.1109/THMS.2020.2988835

Clarke, R. A., & Knake, R. K. (2019). *The fifth domain: Defending our country, our companies, and ourselves in the age of cyber threats*. Penguin Press.

Cooper, G. E., & Harper, R. P. (1969). *The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities*. National Aeronautics and Space Administration.

Coss, R. G. (2017). Drawings of representational images by Upper Paleolithic humans and their absence in Neanderthals reflect historical differences in hunting wary game. *Evolutionary Studies in Imaginative Culture*, *1*(2), 15–38.

Daniels, O. (2021). *Speeding Up the OODA Loop with AI: A Helpful or Limiting Framework?* https://www.japcc.org/essays/speeding-up-the-ooda-loop-with-ai/

Dickmanns, E. D. (1998). Vehicles capable of dynamic vision: a new breed of technical beings? *Artificial Intelligence*, *103*(1-2), 49–76. https://doi.org/10.1016/S0004-3702(98)00071-X

Draper, M., & van Diggelen, J. (2020). *Human Systems Integration for Meaningful Human Control over AI-based systems.* NATO.

Elgezabal, O., & Schumann, H. (2012). Holistic Systems Engineering: Towards a cross-disciplinary standard. In M. Maurer & S.-O. Schulze (Eds.), *Tag des Systems Engineering: Zusammenhänge erkennen und gestalten* (pp. 319–328). Hanser. https://doi.org/10.3139/9783446436039.032

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (Pre-Print)*.

Esfeld, M. (2001). *Holism in Philosophy of Mind and Philosophy of Physics*. *Synthese Library, Studies in Epistemology, Logic, Methodology, and Philosophy of Science: Vol. 298*. Springer Netherlands. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=3106008

Fenn, J., & Time, M. (2007). Understanding Gartner's hype cycles, 2007. *Gartner ID G*, *144727*.

Fischer, J. M., & Ravizza, M. (2000). Precis of Responsibility and Control: A Theory of Moral Responsibility. *Philosophy and Phenomenological Research*, *61*(2), 441. https://doi.org/10.2307/2653660

Flemisch, F., Abbink, D. A., Itoh, M., Pacaux-Lemoine, M.-P., & Weßel, G. (2019a). Joining the blunt and the pointy end of the spear: towards a common framework of joint action, human–machine cooperation, cooperative guidance and control, shared, traded and supervisory control. *Cognition, Technology & Work*, *21*(4), 555–568. https://doi.org/10.1007/s10111-019-00576-1

Flemisch, F., Baltzer, M., Abbink, D., Siebert, L., Diggelen, J., Draper, M., Boardman, M., & Pacaux-Lemoine, M.-P. (2022). Towards a Dynamic Balance between Humans and AI-based Systems: System of system Perspective on Ability, Responsibility, Authority, Autonomy, Meaningful and Effective Control, and Accountability. In L. Siebert & D. Abbink (Eds.), *Handbook on Meaningful Human Control.*

Flemisch, F., Preutenborbeck, M., Baltzer, M., Wasser, J., Kehl, C., Grünwald, R., Pastuszka, H.-M., & Dahlmann, A. (2022). Human Systems Exploration for Ideation and Innovation in Potentially Disruptive Defense and Security Systems. In *Disruption, Ideation and Innovation for Defence and Security* (pp. 79–117). Springer.

Flemisch, F. O., Adams, C. A., Conway S. R., Goodrich K. H., Palmer M. T., & Schutte P. C. (2003). *The H-Metaphor as a guideline for vehicle automation and interaction*. Hampton, Va, USA. NASA Langley Research Center. https://ntrs.nasa.gov/citations/20040031835

Flemisch, F., Abbink, D. A., Itoh, M., Pacaux-Lemoine, M.-P., & Weßel, G. (2019b). Joining the blunt and the pointy end of the spear: towards a common framework of joint action, human-machine cooperation, cooperative guidance and control, shared, traded and supervisory control. *Cognition, Technology & Work*, *21*(4), 555–568.

Flemisch, F., Altendorf, E., Canpolat, Y., Weßel, G., Baltzer, M., Lopez, D., Herzberger, N. D., Voß, G. M. I., Schwalm, M., & Schutte, P. (2017). Uncanny and Unsafe Valley of Assistance and Automation: First Sketch and Application to Vehicle Automation. In C. M. Schlick, S. Duckwitz, F. Flemisch, M. Frenz, S. Kuz, A. Mertens, & S. Mütze-Niewöhner (Eds.), *Advances in Ergonomic Design of Systems, Products and Processes: Proceedings of the Annual Meeting of GfA 2016* (pp. 319–334). Springer. https://doi.org/10.1007/978-3-662-53305-5_23

Flemisch, F., & Baltzer, M. (2022). Are Rider-Horse or Centaurs intelligent Human Systems Integration? First Sketch of reversible and non-reversible human technology/machine/AI Symbiosis. In T. Ahram, W. Karwowski, P. Di Bucchianico, R. Taiar, L. Casarotto, & P. Costa (Eds.), *AHFE International, Intelligent Human Systems Integration (IHSI 2022) Integrating People and Intelligent Systems.* AHFE International. https://doi.org/10.54941/ahfe1001039

Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., & Beller, J. (2012). Towards a Dynamic Balance Between Humans and Automation: Authority, Ability, Responsibility and Control in Shared and Cooperative Control Situations. *Cognition, Technology & Work*, *14*(1), 3–18. https://doi.org/10.1007/s10111-011-0191-6

Flemisch, F., Preutenborbeck, M., Baltzer, M., Wasser, J., Meyer, R., Herzberger, N., Bloch, M., Usai, M., & Lopez, D. (2021). Towards a Balanced Analysis for a More Intelligent Human Systems Integration. In D. Russo, T. Ahram, W. Karwowski, G. Di Bucchianico, & R. Taiar (Eds.), *Advances in Intelligent Systems and Computing. Intelligent Human Systems Integration 2021* (Vol. 1322, pp. 31–37). Springer International Publishing. https://doi.org/10.1007/978-3-030-68017-6_5

Flemisch, F., Schieben, A., Kelsch, J., & Löper, C. (2008). Automation spectrum, inner / outer compatibility and other potentially useful human factors concepts for assistance and automation. In D. de Waard, F. Flemisch, B. Lorenz, H. Oberheid, & K. A. Brookhuis (Eds.), *Human Factors for assistance and automation.* Shaker Publishing. https://elib.dlr.de/57625

Flemisch, F. O. (2022). *Towards a Holistic Understanding of Cognitive Warfare, including Human Factors, Human Systems Integration, Human-Machine Teaming and Human-AI Cooperation*. Report of NATO-STO-RTG 356 "Cognitive Warfare"; in Press.

Foerster, H. von, & Poerksen, B. (2002). *Understanding systems: Conversations on epistemology and ethics*. *International series on systems science and engineering: Vol. 17*. Kluwer; C. Auer-Systeme.

Frank Flemisch, Eugen Altendorf, Yigiterkut Canpolat, Gina Weßel, & Maximilian Schwalm (2016). Arbeiten in komplexen Mensch-Automations-Systemen: Das unheimliche und unsichere Tal der Automation, erste Skizze am Beispiel der Fahrzeugautomatisierung. In GfA (Ed.), *Arbeit in komplexen Systemen - Digital, vernetzt, human?!* (pp. 1–7). GfA-Press. https://www.researchgate.net/profile/frank-flemisch/publication/309012438_arbeiten_in_komplexen_mensch-automations-systemen_das_unheimliche_und_unsichere_tal_der_automation_erste_skizze_am_beispiel_der_fahrzeugautomatisierung

Franklin, G. F., Powell, J. D., & Emami-Naeini, A. (2002). *Feedback control of dynamic systems* (Vol. 4). Prentice hall Upper Saddle River.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gasser, T. M., Arzt, C., Ayoubi, M., Bartels, A., Bürkle, L., Eier, J., Flemisch, F., Häcker, D., Hesse, T., Huber, W., Lotz, C., Maurer, M., Ruth-Schumacher, S., Schwarz, J., & Vogt, W. (2012). *Rechtsfolgen zunehmender Fahrzeugautomatisierung: Gemeinsamer Schlussbericht der Projektgruppe*. *Berichte der Bundesanstalt für Strassenwesen : F, Fahrzeugtechnik: Vol. 83*. Wirtschaftsverlag NW.

Gillespie, A. (2019). *Systems Engineering for Ethical Autonomous Systems*. Institution of Engineering and Technology - IET.

Goodrich, K. H., Flemisch, F. O., Schutte, P. C., & Williams, R. A. (2006). A Design and Interaction Concept for Aircraft with Variable Autonomy: Application of the H-Mode. In IEEE/AIAA (Ed.), *25th Digital Avionics Systems Conference, 2006 IEEE.* IEEE / Institute of Electrical and Electronics Engineers Incorporated. https://elib.dlr.de/50225/

Griffiths, P., & Gillespie, R. B. (2004). Shared control between human and machine: haptic display of automation during manual control of vehicle heading. In *HAPTICS 2004: 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems : proceedings : 27-28 March 2004, Chicago, Illinois, USA* (pp. 358–366). IEEE Computer Society. https://doi.org/10.1109/HAPTIC.2004.1287222

Haberfellner, R., Weck, O. L. de, Fricke, E., & Vössner, S. (2019). *Systems engineering: Fundamentals and applications*. Birkhäuser. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5920010

Hakuli, S., Bruder, R., Flemisch, F. O., Löper, C., Rausch, H., Schreiber, M., & Winner, H. (2009). Kooperative Automation. In H. Winner, S. Hakuli, & G. Wolf (Eds.), *Praxis. Handbuch Fahrerassistenzsysteme: Grundlagen, Komponenten und Systeme für aktive Sicherheit und Komfort : mit 550 Abbildungen und 45 Tabellen* (pp. 647–656). Vieweg + Teubner. https://doi.org/10.1007/978-3-8348-9977-4_43

Hammond, L. M. (2015). *Dieter Zetsche, 1968 Vienna Convention and autonomous cars*. https://www.drivingthenation.com/dieter-zetsche-1968-vienna-convention-and-autonomous-cars/

Hancock, P. A. (1996). Teleology for Technology. In R. Parasuraman & M. Mouloua (Eds.), *Human factors in transportation. Automation and human performance: Theory and applications* (pp. 461–497). CRC Press.

Harcourt, A. H., & Waal, F. B. M. de. (1992). *Coalitions and alliances in humans and other animals*. Oxford University Press.

Helms Mills, J., Thurlow, A., & Mills, A. J. (2010). Making sense of sensemaking: The critical sensemaking approach. *Qualitative Research in Organizations and Management: An International Journal*, *5*(2), 182–195. https://doi.org/10.1108/17465641011068857

Hoc, J. M. (2000). From human-machine interaction to human-machine cooperation. *Ergonomics*, *43*(7), 833–843. https://doi.org/10.1080/001401300409044

Hoc, J.-M., & Lemoine, M.-P. (1998). Cognitive Evaluation of Human-Human and Human-Machine Cooperation Modes in Air Traffic Control. *The International Journal of Aviation Psychology*, *8*(1), 1–32. https://doi.org/10.1207/s15327108ijap0801_1

Hollnagel, E., & Woods, D. D. (1983). Cognitive Systems Engineering: New wine in new bottles. *International Journal of Man-Machine Studies*, *18*(6), 583–600. https://doi.org/10.1016/S0020-7373(83)80034-0

Holzmann, F. (2008). Adaptive cooperation between driver and assistant system. In F. Holzmann (Ed.), *Adaptive Cooperation Between Driver and Assistant System: Improving Road Safety* (pp. 11–19). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74474-0_2

Horowitz, M. C., & Scharre, P. (2015). *Meaningful Human Control in Weapon Systems: A Primer.* Center for a New American Security. http://www.jstor.org/stable/resrep06179

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Ilves, T. H. (2014, January 31). *Rebooting Trust? Freedom vs Security in Cyberspace*. Opening Address at Munich Security Conference, Munich. https://search.proquest.com/openview/6eab2f93756e720dc6608b494985005f/1?pq-origsite=gscholar&cbl=25776

Jones, K. W., Jenkins, L. C., & Ramsey, J. (2005). Holistic systems engineering: physical, informational, and cognitive domains. In IEEE (Ed.), *24th Digital Avionics Systems Conference: Avionics in a Changing Marketplace: Safe and Secure?* (2nd ed., 9.B.6). IEEE. https://doi.org/10.1109/DASC.2005.1563483

Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems*, *19*(6), 91–95.

Lewin, K. (1943). Psychology and the Process of Group Living. *The Journal of Social Psychology*, *17*(1), 113–131. https://doi.org/10.1080/00224545.1943.9712269

Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, *20*(2), 130–141.

Maier, M. W. (1996). Architecting Principles for Systems-of-Systems. *INCOSE International Symposium*, *6*(1), 565–573. https://doi.org/10.1002/j.2334-5837.1996.tb02054.x

Melman, T., Beckers, N., & Abbink, D. (2020). Mitigating undesirable emergent behavior arising between driver and semi-automated vehicle. *ArXiv Preprint ArXiv:2006.16572*.

Merriam-Webster. *control*. https://www.merriam-webster.com/dictionary/control

Miller, C. (Ed.) (2022, March). *Meaningful human control and ethical neglect tolerance: Initial thoughts on how to define, model and measure them*.

National Transportation Safety Board. (2019). *Highway Accident Report NTSB/HAR-19/03: Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018.* Washington, DC. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf

Onken, R., & Schulte, A. (Eds.). (2010). *Studies in Computational Intelligence*. *System-Ergonomic Design of Cognitive Automation*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03135-9

Online Etymology Dictionary. (2022). *Etymology, origin and meaning of university by etymonline*. https://www.etymonline.com/word/university

Pacaux-Lemoine, M.-P. (2020). *Human-Machine Cooperation: Adaptability of shared functions between Humans and Machines - Design and evaluation aspects* [Habilitation à Diriger des Recherches, Université Polytechnique Hauts-de-France, Valenciennes]. hal.archives-ouvertes.fr. https://hal.archives-ouvertes.fr/tel-02959402/

Pacaux-Lemoine, M.-P., & Debernard, S. (2002). Common work space for human–machine cooperation in air traffic control. *Control Engineering Practice*, *10*(5), 571–576. https://doi.org/10.1016/S0967-0661(01)00060-0

Pacaux-Lemoine, M.-P., Debernard, S., Crévits, I., & Millot, P. (1996). Cooperation between Humans and Machines: First Results of an Experiment with a Multi-Level Cooperative Organisation in Air Traffic Control. *Comput. Support. Cooperative Work.*, *5*(2/3), 299–321.

Pacaux-Lemoine, M.-P., & Flemisch, F. (2019). Layers of shared and cooperative control, assistance, and automation. *Cognition, Technology & Work*, *21*(4), 579–591. https://doi.org/10.1007/s10111-018-0537-4

Pacaux-Lemoine, M.-P., & Loiselet, A. (2002). A Common Work Space to Support the Cooperation in the Cockpit ofa Two-Seater Fighter Aircraft. In M. Blay-Fornarino, A.-M. Pinna-Dery, K. Schmidt, & P. Zaraté (Chairs), *Cooperative Systems Design, A Challenge of the Mobility Age,,* Saint-Raphaël. https://www.researchgate.net/publication/221389398_A_Common_Work_Space_to_Support_the_Cooperation_in_the_Cockpit_ofa_Two-Seater_Fighter_Aircraft

Pacaux-Lemoine, M.-P., & Trentesaux, D. (2019). Ethical Risks of Human-Machine Symbiosis in Industry 4.0: Insights from the Human-Machine Cooperation Approach. *IFAC-PapersOnLine*, *52*(19), 19–24. https://doi.org/10.1016/j.ifacol.2019.12.077

Pacaux-Lemoine, M.-P., & Vanderhaegen, F. (2013). Towards Levels of Cooperation. In IEEE (Ed.), *2013 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 291–296). IEEE. https://doi.org/10.1109/SMC.2013.56

Pacaux-Lemoine, M.-P., & Flemisch, F. (2022). Human–Industrial Cyber-Physical System Integration: Design and Evaluation Methods. In O. Cardin (Ed.), *Digitalization and Control of Industrial Cyber-Physical Systems: Concepts, technologies and applications* (Vol. 9, pp. 171–188). John Wiley and Sons Inc. https://doi.org/10.1002/9781119987420.ch10

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans : A Publication of the IEEE Systems, Man, and Cybernetics Society*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Pias, C. (Ed.). (2003/2016). *qu*. *Cybernetics: The Macy Conferences 1946-1953 ; the complete transactions* (First printing). Chicago University Press.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(3), 257–266. https://doi.org/10.1109/TSMC.1983.6313160

Rasmussen, J., & Goodstein, L. P. (1987). Decision support in supervisory control of high-risk industrial systems. *Automatica*, *23*(5), 663–671. https://doi.org/10.1016/0005-1098(87)90064-1

Rausand, M. (2013). *Risk assessment: theory, methods, and applications* (Vol. 115). John Wiley & Sons.

Rizzolatti, G., & Sinigaglia, C. (2008). *Mirrors in the brain: how our minds share actions and emotions*. Oxford University Press.

Ross, W. D. (1924). *Aristotle's Metaphysics: A revised Text with Introduction and Commentary*. Oxford University Press, American Branch. https://www.jstor.org/stable/3288658?seq=1#metadata_info_tab_contents

Rubinsztejn, A. (2018). *Chaos and the Double Pendulum*. https://gereshes.com/2018/11/19/chaos-and-the-double-pendulum/

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*, 15. https://doi.org/10.3389/frobt.2018.00015

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W. W. Norton & Company.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Shea, G. (2019). *2.0 Fundamentals of Systems Engineering*. https://www.nasa.gov/seh/2-fundamentals

Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. John Wiley and Sons Inc.

Siebert, L., & Abbink, D. (Eds.). (2022). *Handbook on Meaningful Human Control.*

Simon, H. A. (1996/2019). *The sciences of the artificial* (3rd ed.). MIT Press.

Sneddon, J. (2017). *Practical Application of Bowtie Analysis: Enhancing Traditional PHA*. https://www.cheminst.ca/wp-content/uploads/2019/04/509-Application-of-Bowtie-CSChE2017.pdf

Snow, C. P., & Collini, S. (1956). *The two cultures*. Cambridge University Press Cambridge.

Spiekermann, S., & Winkler, T. (2020, April 28). *Value-based Engineering for Ethics by Design*. https://arxiv.org/pdf/2004.13676 https://doi.org/10.48550/arXiv.2004.13676

Stefanos, G. B., Efstratios N. Pistikopoulos, & Styliani Avraamidou (2021). A systems engineering framework for the optimization of food supply chains under circular economy considerations. *Science of the Total Environment*, *794*, 148726. https://doi.org/10.1016/j.scitotenv.2021.148726

Stieglitz, S., Mirbabaie, M., & Milde, M. (2018). Social positions and collective sense-making in crisis communication. *International Journal of Human-Computer Interaction*, *34*(4), 328–355.

Strogatz, S. (1995). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, engineering* (3. print). *Studies in nonlinearity*. Addison-Wesley.

Swartz, D. (1997). *Culture and power: The sociology of Pierre Bourdieu*. University of Chicago Press.

Talbert, M. (2019). Moral Responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2019th ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/

Teetor Meyer, M. (2011). *One Man's Vision: The Life of Automotive Pioneer Ralph R. Teetor*. CreateSpace Independent Publishing Platform.

Theodore Roosevelt (1900). *Keep your eyes on the stars and your feet on the ground*, Chicago, Illinois.

Thieme, H. (1997). Lower Palaeolithic hunting spears from Germany. *Nature*, *385*, 807–810. https://doi.org/10.1038/385807a0

Thrun, S. (2006). Winning the darpa grand challenge. In *European Conference on Machine Learning.* Symposium conducted at the meeting of Springer.

Tomasello, M. (2014). *A natural history of human thinking*. *ProQuest Ebook Central*. Harvard University Press. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=3301383

Tsugawa, S. (1994). Vision-based vehicles in Japan: machine vision systems and driving control systems. *IEEE Transactions on Industrial Electronics*, *41*(4), 398–405. https://doi.org/10.1109/41.303790

van den Hof, P. M. J., Scherer, C., & Heuberger, P. S. C. (Eds.). (2009). *Model-Based Control: Bridging Rigorous Theory and Advanced Technology*. Springer.

visual-arts-cork.com. (2022). *Pech-Merle Cave Paintings (c.25,000 BCE).* ENCYCLOPEDIA OF STONE AGE ART. http://www.visual-arts-cork.com/prehistoric/pech-merle-cave-paintings.htm

von Glasersfeld, E. (1984). An Introduction to Radical Constructivism. In P. Watzlawick (Ed.), *The Invented Reality* (pp. 17–40). Norton.

Wanless, M. (2001, 1992). *Ride with your mind: An illustrated masterclass in right brain riding*. Trafalgar Square Pub.

Warren, T. (2012). *Nevada approves regulations for self-driving cars: Nevada becomes the first US state to approve regulations for self-driving cars on its roadways*. https://www.theverge.com/2012/2/17/2804284/google-self-driving-cars-nevada-approval

Weick, K. E. (1974). Middle range theories of social systems. *Behavioral Science*, *19*(6), 357–367. https://doi.org/10.1002/bs.3830190602

Weick, K. E. (1993). The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. *Administrative Science Quarterly*, *38*(4), 628–652. https://doi.org/10.2307/2393339

Weick, K. E. (1995). *Sensemaking in organizations* [Nachdr.]. *Foundations for organizational science*. Sage Publications.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421.

Wiener, N. (1950). Cybernetics. *Bulletin of the American Academy of Arts and Sciences*, *3*(7), 2–4. https://doi.org/10.2307/3822945

Wiener, N. (1954). Cybernetics in history. *Theorizing in Communication: Readings Across Traditions*, 267–273.

Wiener, N. (1961). Cybernetics: Control and Communication in the Animal and the Machine-2nd.

Wiener, N. (2019). *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press.

Zhao Tingyang (2010). The Tianxia System : World Order In A Chinese Utopia. *Global Asia*, *4*(4), 108–112. https://www.semanticscholar.org/paper/The-Tianxia-System-%3A-World-Order-In-A-Chinese-Tingyang/891f9492dab54d95a95940282570866a47e80c77