Robots in the Wild
Report V: Autonomy in Robots Revisited

By Pernille Maja Carlsen, Peter Hommel Østerlund, Jamie Wallace and Cathrine Hasse.

Introduction

This report is the fifth and final report in the *Robots in the Wild* series (see https://robosapiens-eu.tech/robots-in-the-wild/) produced within the RoboSAPIENS project. RoboSAPIENS develops technologies for trustworthy robotic self-adaptation and demonstrates them across four industry-scale use cases related to robot disassembly, robot navigation, ship motion prediction, and human–robot interaction (see https://robosapiens-eu.tech/ "industrial cases"). As the SSH (Social Sciences and Humanities) team in the project, our role is to contribute a humanistic and social perspective to these developments.

The cases and fieldwork presented in this report are not drawn directly from the four industrial RoboSAPIENS use cases. Rather, the case studies in the wild discussed here are included as a supplement because they make visible, in present-day settings, the kinds of challenges that RoboSAPIENS seeks to address: how robots encounter uncertainty, how humans respond to and work with them, and how safety, trust, adaptation, and responsibility are negotiated in practice. The robots examined in this report do not yet have access to RoboSAPIENS technology. Instead, they operate with the available technologies in everyday environments alongside human workers. This works as an empirical lens on the real-world conditions here and now that make the RoboSAPIENS ambition both necessary and difficult. The empirical cases discussed here have all been presented in detail in Reports I–IV; the task of Report V is therefore to synthesise and analyse findings across the series, rather than to introduce new case material.

The broader RoboSAPIENS project is concerned with how robots can become capable of open-ended, trustworthy self-adaptation: that is, how they can autonomously adapt their controllers and configuration settings to unknown changes and variations while remaining safe, robust, and trustworthy[1]. On the project's own terms, autonomy is therefore closely connected to adaptive capacity under changing real-world conditions. At the same time, our first report showed that autonomy is never a simple or self-evident category. In public discourse and at robotics conferences, robots are

1 On RoboSAPIENS homepage (by April 10th 2026) self-adapting robots are defined in this way: "RoboSAPIENS will develop the underlying technologies which enable robots to fully autonomously adapt their controllers and configuration settings to accommodate for unknown changes and variations while ensuring trustworthiness." (https://robosapiens-eu.tech/)

often presented as more autonomous than they are in practice (see Report I https://robosapiens-eu.tech/wp-content/uploads/2026/02/Report-I-Robot-Autonomy-and-Conferences_AA.pdf). Their movements, interfaces, and staged performances invite observers to read autonomy from outward behaviour. By contrast, engineering approaches tend to define autonomy operationally, as something technical in feedback loops that monitor, analyse, plan, and execute, and in RoboSAPIENS more specifically through the new MAPLE-K framework and related trustworthiness mechanisms (Larsen et al. 2024). Report I argued that this engineering understanding is important, but not sufficient on its own. It proposed instead that autonomy must also be understood as lived, relational, and situated: something that emerges in practice through ongoing interaction between robots, humans, and environments (Report I, pp. 21–23, 26).

Across the fieldwork reported in Reports I–IV, autonomy does not appear as a stable property that a machine simply possesses once the right technical threshold has been reached. Instead, it appears as something continuously achieved, adjusted, supported, interrupted, and sometimes undone in practice. Robots do not operate alone in any simple sense. Their functioning depends on human oversight, organisational routines, safety procedures, infrastructural arrangements, environmental conditions, and situated judgement. In this sense, autonomy in the wild is not stand-alone autonomy, but distributed autonomy.

The comparative findings across the reports make this especially clear. Report I showed that robot autonomy is often imagined and publicly staged as more complete than it is in practice. Report II showed that public spaces expose the fragility of such autonomy, because ordinary human unpredictability repeatedly exceeds what robots can smoothly handle. Report III showed that in controlled environments autonomy is not simply achieved by better machines, but by routines, training, workarounds, and distributed responsibility that stabilise limited robotic capacities. Report IV showed that in dynamic environments such as agriculture and maritime navigation, these compensations reach their limits, because uncertainty, changing conditions, and the need for practical judgment remain irreducible. What these studies show is not simply that present robots are “not yet autonomous,” but that real-world autonomy is always socio-technical: it is shaped through relations among machine capabilities, human practices, institutional arrangements, and material environments.

More specifically the ambition of the RoboSAPIENS project is to enable robots to modify their behaviour in response to changing environmental conditions and interactions with humans, while maintaining guarantees of safety, efficiency, and trustworthiness. Technically, the project builds upon the established MAPE-K architecture, extending it into a novel MAPLE-K architecture (Larsen et al. 2016, Larsen et al. 2024). This extension introduces two significant mechanisms intended to strengthen runtime assurance and to embed safety more explicitly within the adaptive loop. The first of these is the introduction of a “Legitimate” phase, which functions as a validation and verification step for any proposed adaptation. The second addition is a run-time trustworthiness checker embedded within the robot’s controller. This mechanism is designed to reduce uncertainty during operation by continuously evaluating whether ongoing and planned behaviours meet defined criteria for robustness and safety. However, its ultimate success will depend on the extent to which its technical mechanisms remain responsive to the complexities of real-world use. Some of these challenges are discussed in a chapter on design and user issues at the end of this report.

From this perspective, the contribution of *Robots in the Wild* to RoboSAPIENS is to reframe self-adaptive autonomy from a purely technical question to a broader understanding of how self-adaptive autonomy has implications for Human-Robot-Interaction (HRI). This is where the five recurring analytical themes developed across the reports become important: Safety, Trust, Anomalies, Sim2Real and Sabotage/Tinkering. These are not secondary social issues added onto an otherwise technical core. They are the very sites where autonomous adaptation becomes possible, limited, contested, or redistributed in real-world use. Safety shows that adaptation should remain accountable in practice, not only in design. Trust shows that self-adaptive autonomy depends on how systems are experienced, interpreted, and entrusted by those who work and live with them. Anomalies reveal where adaptation is needed, where existing models break down, and where the practical limits of autonomy become visible. Sim2Real shows why adaptation cannot rely on models alone but should contend with the persistent gap between simulation and lived environments. Sabotage and Tinkering show that human interventions are not merely interference, but often crucial forms of feedback and situated design knowledge. Report I already identified these as recurring themes linking engineering concerns with lived realities, and Report V brings together the findings from the earlier reports in order to show how autonomy, as encountered in the wild, is co-produced through technical systems and human practice alike.

The central argument of this report is therefore that robot autonomy should be revisited not as independence from humans, but as a distributed and negotiated accomplishment. If RoboSAPIENS seeks robots that can adapt under unknown change while remaining trustworthy, then the question is not only how robots adapt technically, but also how such adaptation is made legible, safe, accountable, and workable for the people who live and work with these systems. Our purpose is not to dismiss the RoboSAPIENS ambition to create self-adaptive and trustworthy robotic systems. Rather, it is to extend and ground that ambition empirically as well as open for new research questions.

This report adopts a speculative approach to the RoboSAPIENS technology yet to be developed, exploring how new technical understandings of autonomy may translate into the complex and situational realities in which robots of the future are expected to operate in the wild. As such, this work should be understood as ongoing and exploratory. While we do not propose concrete actions for the development of accountable, self-adaptive robots, given that we are not engineers, we aim instead to reflect on the broader implications of such technologies in real-world contexts. Importantly, our speculations are not purely hypothetical. They are grounded in multiple field studies on HRI, as outlined in our earlier reports. The questions that emerge from this inquiry into accountable, self-adaptive robots are therefore not endpoints, but contributions to an evolving process of understanding and design.

This process is inherently non-linear and iterative, as evidenced by fieldwork documenting how humans continuously negotiate and adapt to robots in everyday situations. Such findings reinforce the idea that both the technical and social dimensions of these systems are highly complex and deeply intertwined. In the following we present our findings in five chapters (following the order of Safety, Trust, Anomalies, Sim2Real, and Sabotage/Tinkering) ending with an overall conclusion drawing findings together.

Contents

Overview of the main findings

Safety	Trust	Anomalies	Sim2Real	Sabotage/Tinkering
Safety as promised • Presentation in media and at conferences **Safety as formalised** • Risk assessments, standards, design and programming **Safety as enacted** • Trade-offs, techno-regulation, accountability and perceived safety	**Trust about the robot** • Formalised trust • Expectations and apperances **Trust in the robot** • Opacity, technical knowledge, responsibility, control, physical trust • Three layers of trust in practice ◦ Physical ◦ Functional ◦ Informational	**Valence** • Whether anomalies are experienced as disruptive (negative) or productive (positive) in practice **Robot-perceived anomalies** • Negative: Leads to stops or degraded performance • Positive: Expands what the system can treat as "normal" **Human-perceived anomalies** • Negative: Creates extra work • Positive: Builds local know-how	**Simulation reduces complexity but cannot keep pace with lived environments** **Simulation is useful because it prepares, reveals and diagnoses** **Humans continiously align model and world** **Simulation embed selective worlds**	**Sabotage and tinkering exist on a dynamic spectrum** **Learning the system: testing, probing, and informal skill-building** **Making the system workable: Workaround, correction and risky compensation** **Resistance, refusal and legitimacy crisis**

[Overview of the main findings presented in the chapters below]


ROBOSAPIENS
ROBOSAPIENS

Safety

"In order to secure safety when designing collaboration with robots, there are different safety measures available, such as infrared beams that work as fences. In connection with this, Nick from ElectroFlight says: "Fences are not put up so that the robot cannot get out. They are put up so that employees cannot get in." (Report III, p. 7). Safety is not just about the humans but also robots that must be kept safe from human tinkering. The risk assessment of the robot is in fact also to protect the robot from human interference. Nick elaborates that this is because humans do not take the safety measures seriously and think they can just go in to the robot or move the fence. When we talk about the different standards and regulations that exist in relation to safety, Nick explains that as a manufacturer he has the responsibility that no one gets hurt with his products, and therefore they make sure to follow the rules. However, he immediately adds that he cannot control what the customers then choose to do afterwards. He compares this to a car manufacturer and seat belts: "What they then come up with afterwards, just like if you drive your car and say, well, we can't be bothered to use the seat belt. You can't blame the manufacturer for that. It was there, it worked, but you're not using it." (Report III, p. 7). Here Nick points out the complexity of designing safety when working with humans, since they (humans) can be unpredictable and unreliable."

[Introductory vignette drawing on Report III illustrating several aspects of safety dealt with below.]

Introduction

Safety is one of the most central and most valued concepts in RoboSAPIENS. In the project's own terms, self-adaptive robots must be able to respond to unknown changes and variations while remaining trustworthy, and such adaptation cannot be boundless: it must take place while maintaining or even increasing expected performance and staying at least as safe and robust as before. Safety is therefore not an auxiliary concern within RoboSAPIENS, but one of the core conditions of adaptive autonomy itself. In the project material, this ambition is further tied to formal verification, validation, and the use of safety engineering techniques able to ensure that actions remain within acceptable bounds during and after self-adaptation.

This chapter does not challenge the importance of that formal ambition. On the contrary, our starting point is that formal safety is indispensable. Risk assessments, standards, regulations, safety envelopes, emergency stops, collision limits, environmental sensors, and verification procedures are all necessary achievements. They are what make robots marketable, certifiable, and, in many cases, usable at all. In this sense, safety is often treated in engineering as something that is specified, designed, and tested in advance. As this safety chapter will show, RoboSAPIENS itself adopts such a definition when safety is described as preventing accidents and damage through validation and verification against relevant standards.

However, our fieldwork across restaurants, warehouses, hospitals, farms, conferences, and maritime settings shows that formal safety is not the same thing as safety in use. Once robots move outside controlled demonstrations and into

shared, ordinary environments, safety does not appear simply as a property located in the machine or guaranteed by prior design. It appears instead as something that should be continuously achieved in practice through relations among technical systems, humans, environments, infrastructures, routines, and institutions. Safety, in this sense, is not only designed into robots; it is also enacted around them, interpreted through them, and revised because of them. This broader point is consistent with the report's general argument that autonomy in the wild is distributed rather than self-contained, and that adaptive systems should be made legible, accountable, and workable for the people who live and work with them.

This section therefore treats safety as a socio-technical process that moves across three linked aspects. First, safety is promised: it is imagined, staged, and publicly presented in particular ways before deployment Second, safety is formalised: it is translated into standards, risk assessments, technical safeguards, and design and verification procedures. Third, safety is enacted: it is achieved in practice through local routines, embodied judgments, environmental adjustments, negotiations and distributed responsibility and then a final synthetic section where we clarify our empirical understanding of safety. In this fourth section we take into account how incidents, workarounds, perceived risks, and new situations feed back into redesign, retraining, new rules, and new expectations. This framing allows us to connect the engineers' highly valued notion of safety with our ethnographic finding that safety is always also "in the making."

The empirical material in this chapter shows this repeatedly. Safety is formalised through standards and risk assessments, but these do not fully determine what happens once the robot enters a real setting. Workers develop local routines around blind spots, speed limits, stop buttons, fences, sensor failures, and false alarms. They learn when a robot is "safe enough," when it is too sensitive to be practical, when rules can be stretched, and when intervention is required. They also learn when formal safety features create new problems of efficiency, workflow, or trust. Safety thus becomes a site of ongoing learning, not only for machines, but for humans and organisations learning how to work with machines under changing conditions. What begins in design as a technical requirement becomes in everyday use a cultural-material learning process in which safety is negotiated through breakdowns, workarounds, judgments, and trade-offs.

This is why the central question of this section is not whether safety is either formal or negotiated. The point is rather that these two are inseparable. Formal safety provides the necessary starting framework: it defines acceptable behaviours, codifies responsibilities, and stabilises expectations. But once the robot is deployed, safety must also be enacted and sustained in practice by people who encounter situations that standards cannot exhaust in advance. In public-space settings, this includes navigating inattentive customers, children, rearranged furniture, and congested paths (Report II). In organisational settings, it includes balancing safety against productivity, usability, and reliability (Report III). In dynamic environments such as farming and maritime navigation, it includes dealing with weather, wildlife, misreadings, uncertain data, and situations in which rule-following alone is not enough (Report IV). Across all of these

settings, safety remains real and important, but it no longer appears as fixed. It appears as a moving settlement between specification and practice.

Seen from this perspective, the contribution of the section to RoboSAPIENS does not replace formal safety engineering with a social account of safety. Rather, it shows what formal safety has to contend with if robots are to remain safe under real and changing conditions. This is especially important for self-adaptive systems. If RoboSAPIENS aims to develop robots that adapt to unknown changes during their lifetime, then safety cannot be treated only as a design-time property checked once and for all. It must also be understood as something that remains legible, interpretable, and revisable in deployment. Humans will still have to notice problems, understand system behaviour, decide when to trust or intervene, and remain accountable when things go wrong. In that sense, our ethnographic material extends the formal ambition of RoboSAPIENS by showing the practical conditions under which safety can actually be sustained in the wild.

The chapter begins, therefore, not with enacted safety itself, but with safety as promised. Before safety is formalised in standards or negotiated in practice, it is already imagined in media, demonstrations, and public expectations. These promised versions of safety matter because they shape what customers, users, workers, and audiences think robots can do, how autonomous they appear to be, and how safe human-robot interaction (HRI) is expected to feel. “Imagined safety” is thus not a dimension of enacted safety in its own right. It is a framing contrast. It reveals the gap between publicly staged safety and the much more structured, cautious, and work-intensive forms of safety that appear in real deployment.

1. Safety as promised

Before safety is encountered in risk assessments, standards, or everyday work, it is often encountered as a promise. Robots circulate in science fiction, media coverage, promotional materials, conferences, fairs, and staged demonstrations as figures of smooth, reliable, and often highly autonomous operation. In such settings, safety is rarely presented as a difficult socio-technical accomplishment. Instead, it is implied through the robot’s appearance, movements, and public framing. Some robots are imagined as harmless, friendly companions; others as powerful threats. But in both cases, safety tends to be represented as though it were an intrinsic quality of the machine itself. The result is that robots appear either naturally safe or naturally dangerous, rather than as technologies whose safety depends on design choices, environmental controls, formal procedures, and human practices.

This matters because these public imaginaries shape expectations long before robots are encountered in actual workplaces or public settings. As fieldwork from Report I and the current section show, conferences and demonstrations often present robots in a curated manner that emphasises capability, smoothness, and future potential. Demonstrations function partly as marketing performances. Safety is therefore not always shown under the conditions in which it would be required in actual deployment. Industrial robots may be displayed without the full fencing or separation that would normally surround them in production settings. Safety thresholds may be adjusted so that demonstrations run more fluidly and appear less interrupted. In one observed case, an operator explained that he had lowered the sensitivity of a robot’s safety settings because otherwise it would stop too often during the demonstration (Report I). Such moments

do not necessarily create direct danger, but they do reveal how public presentations can blur the line between realistic operation and staged performance.

What is produced in such settings is not safety in the practical sense that later concerns manufacturers, operators, regulators, or workers. It is better understood as promised safety: a public impression that robots can move among humans smoothly and safely, with little visible friction. This promised safety is reinforced by wider cultural representations. Science fiction has long imagined robots as autonomous beings operating naturally among people, whether as harmless helpers like C-3PO or as dangerous antagonists like those in *I, Robot*. Although such representations differ in tone, they share the same underlying simplification: they suggest that safety is already solved once autonomy has been achieved (Report I, p. 13). The more mundane reality observed in the fieldwork is very different. In practice, safety is not built into the robot in any simple or final way. It is achieved through risk assessments, environmental restrictions, standards, fences, sensors, speed limits, human oversight, and local adaptations.

The gap between the promised safety and the practical safety also appears in customers' and users' expectations. As the interview with Nick from ElectroFlight (see introductory vignette) makes clear, clients often arrive with inflated ideas of what robots, especially collaborative robots, can safely do. They may expect cobots to combine heavy, continuous industrial performance with close, flexible, and inherently safe collaboration with humans. But such expectations collide with practical limitations. Cobots can be safer for closer human interaction, yet they are not necessarily suited for prolonged high-intensity industrial use without wear, constraint, or reduced performance. Industrial robots can handle those demands more robustly, but usually only under much stricter safety arrangements such as physical separation. In that sense, promised safety creates an image of compatibility between autonomy, productivity, and seamless human proximity that real systems often cannot sustain.

This is why "safety as promised" is a framing device in this section rather than an analytical dimension of safety itself. Its function is not to explain how safety is enacted in practice, but to show the contrast against which the rest of the section unfolds. It helps explain why users, customers, and audiences may initially approach robots with assumptions that do not match the realities of deployment. It also helps explain why later chapters on trust, anomalies, and sabotage encounter disappointment, irritation, or resistance: expectations have already been shaped by a public world in which robots appear safer, smoother, and more autonomous than they typically are in ordinary use. Safety as promised is therefore analytically important not because it is itself a mode of enacted safety, but because it frames the transition from imagined robotics to robotics in the wild.

From the perspective of RoboSAPIENS, this contrast could be especially significant. If the project seeks to develop self-adaptive robots that remain safe and trustworthy under unknown changes, then it should also take seriously the gap between public imagination and practical deployment. The issue is not simply that robots are oversold, but that promised safety can obscure the amount of formal work, situated judgment, and continuing human involvement that real safety requires. By beginning with safety as promised, the chapter therefore prepares the reader to see why the subsequent discussion of formalised, enacted, and empirical safety is necessary. It shifts the question from how robots appear to be safe, to how safety is actually made possible under real conditions.

2. Safety as formalised

If “safety as promised” concerns how robots are publicly imagined before deployment, then “safety as formalised” concerns how safety is specified before the robot is expected to operate in the world. This is the domain in which RoboSAPIENS and safety engineering begin, and where safety is treated as something that is translated into requirements, criteria, safeguards, procedures, and verifiable limits before a system is put into use. It is the domain of pre-deployment anticipation: of asking in advance what could go wrong, how harm might occur, what counts as acceptable risk, and how such risk can be reduced through design, compliance, and verification. In this sense, formalised safety does not stand in opposition to the ethnographic account developed in this report. It is the necessary starting point. Without formalised safety, there would be no safety envelope at all for later practice to work within.

This is also how safety is approached within RoboSAPIENS itself. In the working definitions of the project, safety is described as the robot’s ability to prevent accidents, injuries to humans and living creatures, and damage to itself and its environment. More specifically, safety is tied to the architecture of the system through the Legitimize node of the MAPLE-K loop, which examines through validation and verification whether every planned action complies with the relevant safety requirements. In this formulation, safety is not simply a desirable outcome but a condition that must be checked before action is executed. The broader trustworthiness definitions reinforce this emphasis by linking trustworthiness to robustness in dynamic and uncertain environments, to validation and verification processes, and to governance and regulation of design and operation. Formal safety therefore appears here as one of the core normative and technical foundations of trustworthy self-adaptation.

Seen from this perspective, formalised safety involves at least three closely connected activities. First, there is the translation of safety into standards and regulations. Second, there is risk assessment: the attempt to identify hazards in advance and classify them systematically. Third, there is the material and programmatic embedding of safety into the robot and its environment through sensors, speed limits, interlocks, barriers, interfaces, and control logic. Together, these activities turn safety from a general aspiration into something specified and operational. They are what make it possible for robots to be certified, sold, deployed, and trusted at all. At the same time, our empirical material shows that this layer is never fully neutral or complete. Even here, before the robot enters full practice, formalisation involves interpretation, prioritisation, and judgment.

Risk assessment, standards, and regulations

The clearest expression of safety as formalised appears in the textual and institutional frameworks through which safety is specified before robots enter use. In our material, this formal layer is shaped above all by three partly overlapping regimes: first, ISO standards; second, the EU machinery law; and third, the more recent EU framework around trustworthy and regulated AI. These are not the same kind of texts, nor do they operate in the same way. However, together they define much of the legal and technical landscape within which robots are made, assessed, marketed, and governed. They turn safety from a general aspiration into something that can be written down, checked, documented, and, at least in principle, demonstrated before a system is deployed. In this sense, they form part of the infrastructure that allows a robot to count as acceptably safe in both the empirical fields we studied and in RoboSAPIENS itself.

The first of these texts, ISO standards, are not laws in themselves but internationally agreed standards developed by experts. ISO describes them as "the best way of doing something," ((https://www.iso.org/standards.html) and explains that they are produced through a consensus process involving manufacturers, users, regulators, customers, and other stakeholders. In that sense, ISO standards are technical formulations of what counts as state-of-the-art practice at a given moment. They provide shared methods, concepts, and expectations that can be used across sectors, including health and safety, quality management, and information security. For robotics, their importance lies in precisely this standardising role: they offer technical benchmarks and common vocabularies through which safe design, risk reduction, testing, and conformity can be organised. Their authority is practical and professional rather than legislative, but because they are widely recognised and often linked to conformity procedures, they become highly consequential in the field. They are a way of stabilising expectations about what a safe machine should be and how that safety should be demonstrated.

In practice, safety in the field of robotics is supported through a set of different standards. For example, general risk assessment principles are found in ISO 12100 and guide the identification and mitigation of hazards whilst robot-specific standards such as ISO 10218 and ISO/TS 15066 define requirements for safe operation and human–robot interaction. Software and AI-oriented standards like ISO/IEC 23053 help structure development, validation, and ongoing system management. ISO 10218 for example, specifies requirements for the safety of industrial robot applications and robot cells. It specifies how in cases involving human proximity then safety is achieved through environmental monitoring to control the robot's behaviour. A key requirement here is that robots must implement protective measures to prevent human contact with hazardous motion, such as "Safety-rated monitored stop: the robot automatically stops when a human enters a defined workspace" and "Presence-sensing safeguarding (e.g. light curtains, scanners)" (International Organization for Standardization 2025).

Where ISO standards are expert standards, the EU Machinery Directive is binding product law (https://www.vdma.eu/en/machinery-directive). In our empirical material it appears as one of the central legal frameworks governing robotics, and that is accurate for the settings studied. The Directive lays down the essential health and safety requirements machinery must meet in order to be placed on the EU market, while also promoting the free movement of machinery within the internal market. Importantly, it does not work alone. It follows the familiar European approach in which mandatory legal requirements are combined with voluntary harmonised standards that reflect the state of the art. This distinction matters. The law from 2006 defines the essential requirements and conformity routes; the standards provide recognised technical means of meeting them (Official Journal of the European Union, 2006). As of April 2026, this remains the operative framework for machinery placed on the market before 20 January 2027, even though the newer Machinery Regulation has already been adopted and will apply mandatorily from that date onward. The new Regulation keeps the same basic logic but adds greater legal uniformity, includes provisions for AI-powered safety functions, and explicitly integrates cyber-safety for compliance-relevant software and safety control systems. This makes the machinery framework relevant for RoboSAPIENS, since adaptive and AI-enabled robotic systems increasingly sit at the boundary between conventional machinery law and newer digital regulation.

The third text is the newest and, in some ways, the most unsettled. In this, the AI-specific governance enters primarily through the Ethics Guidelines for Trustworthy AI, published by the European Commission's High-Level Expert Group on AI in 2019 (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai).

These guidelines are not binding law, but they have been highly influential, including in RoboSAPIENS' own working definitions. They formulate trustworthy AI through seven requirements, including human agency and oversight, technical robustness and safety, transparency, accountability, privacy and data governance, fairness, and societal and environmental well-being. In other words, they broaden the meaning of safety beyond bodily protection alone and connect it to robustness, oversight, explainability, and governance. This broader language is echoed in the RoboSAPIENS trustworthiness definitions, where safe operation is linked not only to preventing harm, but also to validation and verification, accountability, understandability, and governance. At the same time, the legal situation has changed since 2019. In the current EU landscape, the more binding AI-specific framework is the AI Act, which entered into force on 1 August 2024 and introduces a risk-based system of obligations, including for high-risk AI systems. For such systems, the Act requires risk assessment and mitigation, documentation, logging, human oversight, robustness, and traceability. The 2019 guidelines therefore remain important as a normative and conceptual reference point, while the AI Act increasingly provides the harder regulatory edge.

Taken together, these three textual regimes do different but connected kinds of work. ISO standards articulate expert consensus and technical best practice. EU machinery law defines binding essential health and safety requirements and the conformity framework for market entry. The AI guidelines and now the AI Act extend the regulatory horizon into questions of adaptive, data-driven, and high-risk AI, bringing safety into closer relation with transparency, accountability, human oversight, and system robustness. Their differences are therefore significant. One is primarily standard-setting, one is product-safety law, and one is a broader AI governance framework moving from soft-law ethics to binding regulation. But their connections are just as important. In practice, they are layered together: the law points toward standards, standard operationalised legal expectations, and AI governance increasingly reshapes what will count as acceptable safety for systems that learn, adapt, or rely on complex software. This is also why they matter so much for RoboSAPIENS. The project's own definition of safety in MAPLE-K (see introduction), especially the role given to the Legitimize node in checking planned action against relevant safety and health requirements, sits precisely at the intersection of these layers (Larsen et al. 2024). Safety is understood not only as a hoped-for outcome, but as something to be checked against standards, regulations, and wider trustworthiness expectations before action is allowed to proceed.

Yet one of the main lessons of our material is that these texts do not by themselves make safety automatic. They codify, classify, and anticipate, but they do not remove interpretation. They tell us that harm must be prevented, that essential requirements must be met, and that trustworthy or high-risk AI must be robust, documented, and overseen. But they do not tell us, by themselves, exactly how a concrete risk should be judged in a given case, how competing harms should be weighted, or how emerging technologies should be handled when standards lag behind practice. This becomes particularly visible once one moves from these general texts to the actual procedures of risk assessment. It is here that

the formal layer comes closest to practice: in the matrices, categories, and judgments through which potential harms are identified, ranked, and mitigated. For that reason, the next step in our analysis turns from these overarching legal and standard-setting texts to the more concrete practice of risk assessment as we encountered it through the safety course and in our empirical fieldwork.

Negotiating standards

From an industry perspective, thorough risk assessment is expected to take place before deployment. As Nick from ElectroFlight explained, risks should not emerge "after the fact," as the system is designed to anticipate and mitigate them in advance (Report III, p. 8). One example is the use of a "safety zone" around robots, created through infrared sensors that detect movement. If something enters this zone, the robot stops immediately as shown in the introductory vignette. Such measures illustrate how safety is often built into both the robot and its surroundings.

However safety regulations also seem to be constantly interpreted and negotiated even by experts presenting the official regulations to people working in practice. This became clear as we, as part of our fieldwork, enrolled in safety courses to explore how the regulations stipulated in the ISO-standards, the Machine Directive and AI-regulations were presented to robot practitioners in the field.

Through online searching in several European countries, we for instance found a site offering a free safety course, which we completed. Afterward, we contacted the company that provided the course and spoke with one of the safety experts (here named Rosa) involved in creating it. Insights from the safety course further demonstrate how risk assessments are conducted in practice. A general risk assessment tool is the Risk Assessment Matrix (figure 1), in which you identify potential hazards and evaluate them based on severity, frequency, likelihood, and the possibility of avoidance. These factors were combined to categorize risks as low, medium, or high, with priority given to those posing the greatest potential harm.

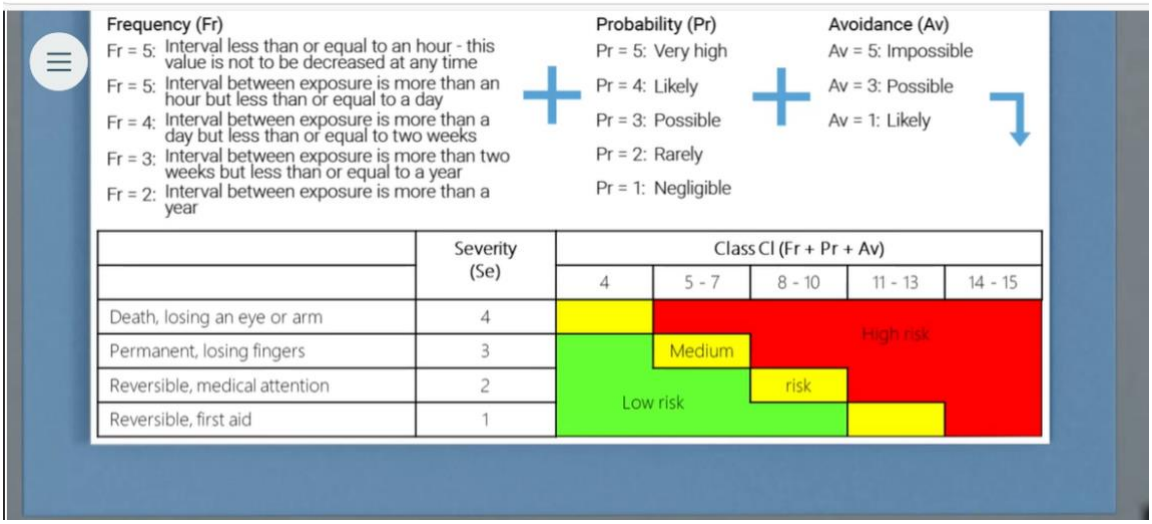


Figure 1: Risk Assessment Matrix

However, this process also revealed an important limitation: risk assessment is not entirely objective. Evaluating the severity of a potential injury can vary between individuals, as perceptions of what constitutes "serious harm" differ. For example, a scenario involving a robot arm striking a person's face might be categorized differently depending on how severe the outcome is believed to be. In the highest risk category number 4 we can for instance see that the matrix is grouping 'death' with 'loosing and eye or an arm'. Such standardized overall categories, grouping "death" with the loss of a limb or eye, can oversimplify complex situations by forcing risks that are very diverse in practice into fixed classifications.

This complication was also acknowledged by Rosa who emphasized that while individual assessments may vary, all risk evaluations are guided by a shared principle: avoiding harm to humans (Report III, p. 23). This requires prioritization, and as some risks are easier to mitigate than others, attention must be directed toward those with the most severe potential consequences. Risk assessment, therefore, involves not only identifying dangers but also navigating how and when to address them.

Rosa also noted that standards can feel restrictive for those working with robots. As she put it, "Anybody who's involved with using robots doesn't want the safety standards. Yeah. They say, I just want to do what I want to do" (Report III, p. 24). She then described MakeBots approach to safety: "We can't be enabling stupid" (Report III, p. 24). Their robots must be safe because users may not voluntarily follow all safety measures, and designers cannot assume compliance.

Documentation plays a crucial role in this process. As explained by Henry from FarmGO, maintaining detailed records of risk assessments is essential, not only for certification but also for accountability (Report IV, p. 8). In the event of an accident, companies must be able to demonstrate that they have identified potential hazards and implemented appropriate mitigation measures. Proper documentation can determine whether an incident is considered negligence or an unavoidable risk.

At the same time, standards can shape the direction of innovation. In the case of FarmGO, their robots are not yet capable of distinguishing between humans and trees. While such functionality is technically feasible, it has not been prioritized because current standards do not require it, particularly as they have yet to fully incorporate AI and machine learning. As a result, development is postponed until such capabilities become part of formal requirements. This reflects a broader dynamic in robotics development: innovation is guided not only by technological possibilities but also by regulatory frameworks. Standards act as roadmaps, indicating which features are necessary for compliance and market entry.

Rosa explained that these standards have a "state of the art" approach, which is to ensure that they are grounded in real-world applications and practical needs, rather than speculative or theoretical technologies (Report III, p. 23). In other words, standards are not meant to predict future innovations but to formalize best practices for technologies that are already being produced, sold, or widely considered for deployment. This approach reinforces the earlier observation that ISO standards are developed alongside practice They organise practices based on existing products, meaning that innovation can sometimes outpace regulation. As a result, developers must balance the potential of new technologies with the current absence of formal safety standards, often making judgment calls on risk mitigation and design decisions before standards catch up.

Safety and security

In one of our discussions with engineers, a distinction was drawn between safety and security. To illustrate it, they sketched a robot at the centre of a diagram with arrows moving both outward from the robot and inward toward it. The outward arrows represented safety, understood as the effects of the robot on its surroundings: whether the robot might injure people, damage objects, or otherwise create harmful consequences in its environment. The inward arrows represented security, understood as the effects of the surroundings on the robot: whether the system might be exposed to hacking, interference, manipulation, or other external threats that could compromise its functioning. The distinction is a useful one because it reflects an important strand of formal safety engineering, in which different kinds of vulnerability are separated in order to make them governable. One concerns how the machine acts on the world; the other concerns how the world acts on the machine.

As an engineering distinction, this separation has clear value. It helps specify different kinds of requirements, different forms of testing, and different domains of responsibility. In the formal frameworks discussed above, safety is primarily associated with avoiding harm through design, validation, verification, and compliance with essential health and safety requirements. Security, by contrast, concerns the protection of systems, data, and operations against unauthorised access, cyberattack, or malicious interference. In that sense, the distinction resonates with the broader regulatory development outlined earlier in the chapter, where newer AI- and software-related frameworks increasingly bring cyber-vulnerability into conversation with more established machinery safety requirements. Even within RoboSAPIENS' own working definitions, this separation is visible: safety is likely treated as a core part of trustworthiness and is integrated into the MAPLE-K architecture, while security is acknowledged as equally relevant but remains less central to the project's trustworthiness checkers. The distinction therefore belongs to the formal layer of safety thinking.

At the same time, our empirical material suggests that the two dimensions are not easily separable once robots move into real environments. The distinction may be analytically useful, but in practice safety and security continuously shape one another. A security breach may become a safety issue if it affects the robot's behaviour in ways that endanger people or equipment. Conversely, many of the safety measures described in our fieldwork also protect the robot from human interference. This is already visible in the introductory vignette from ElectroFlight, where fences, infrared beams, and restricted zones are presented not only as ways of protecting workers from the robot, but also as ways of protecting the robot from workers who move barriers, enter restricted areas, or otherwise disrupt its safe functioning. As Nick puts it, the fence is not there primarily so the robot cannot get out, but so that employees cannot get in. The same physical boundary is therefore both a safety device and a protection against external interference.

This overlap is important because it reinforces a broader argument running throughout the section: safety cannot be understood only as a predefined property embedded inside the machine. Even in the formal register, it already depends on relations between robots and environments, between systems and users, and between anticipated behaviour and possible disruption. The engineers' distinction is therefore helpful not because it gives us two fully separate categories, but because it makes this relationality visible. It draws attention to the fact that robotic systems are never isolated. They operate in environments that can be affected by them, but that also act back on them through misuse, interference, hacking, uncertainty, or simply the unpredictability of human practice. For that reason, safety is not exhausted by outward control and security is not exhausted by inward protection. In practice, the two operate together and are continually reconfigured through the same socio-technical relations.

Seen in this light, the distinction between safety and security provides a final clarification within the formal layer before the chapter turns more fully toward design and practice. It shows that even the most technical language of safety already presupposes a world of interaction rather than isolation. What the next sections explore in greater detail is how this formal ambition is then carried into the material design of robots and, later still, how it is tested, stretched, and reworked in everyday use.

Designing and Programming Safety

When developing robots and robotic systems, engineers and designers employ a range of techniques to ensure safety is embedded into the system prior to its deployment in real-world contexts. These approaches are often guided by existing regulations and standards, which shape how safety is conceptualized and implemented. Safety measures can be integrated at multiple levels, either through the robot's software, by programming constraints and control mechanisms into its operating system, or through its physical design, such as structural features that limit potential harm. The following examples illustrate how safety is designed and operationalized in practice, based on insights from our empirical data.

From our conversations with engineers and developers, it became evident that a key factor in ensuring robot safety is the amount and quality of data the robot can gather. The more advanced the sensors and cameras, the better the robot can understand its environment and act safely (Report I, p. 21). However, equipping a robot with an excessive number of sensors or cameras is not always feasible due to cost and design constraints. There are also ethical concerns around

surveillance, though these are outside the scope of this report, which focuses primarily on safety. Nonetheless, it is evident that more data generally allows for safer operation.

At MedCare, the staff explained that robots can have blind spots (Report III, p. 49). Approaching the robot from certain angles can make it unable to detect a person, which could lead to accidents if someone is in a vulnerable position. Laura and Andy shared an incident in which an employee was halfway run over due to such a blind spot. Importantly, the story was framed not as fear of the robot, but as practical knowledge of its limitations. By understanding where the robot's sensors and cameras cannot reach, humans can act carefully around it without requiring excessive modifications to the robot itself.

The company has anticipated some of these limitations, recognizing that it is impossible to completely eliminate risk when humans are involved. This is one reason employees are prohibited from entering certain robot zones. It is also why they are not allowed to alter the robots' programming: the manufacturer must remain responsible for the safe operation of the machine. This approach emphasizes that safety is not only built into the robot, but is also a negotiated process between human behaviour, awareness, and technological design.

The robots discussed in Report II were all operating in highly public spaces, such as restaurants, stores, and museums. One safety measure they all shared was operating at slow speeds. Moving slowly allows the robots to detect unexpected obstacles or people in their path and have enough time to stop or adjust their movements.

This precaution is particularly important in environments with many people who may have little or no experience interacting with robots. Visitors might walk in front of the robot or behave unpredictably, so the robot must rely on its sensors to avoid collisions, making speed a critical factor in safety.

At one restaurant, a waiter mentioned that the robot occasionally bumps into staff (without causing injury), and this demonstrates that even in controlled environments, collisions can occur. The presence of children, who tend to run around and are less aware of their surroundings, further emphasises the need for slow movement as a primary safety measure.

Similarly, at MUSEA and SHOPFLOOR, staff highlighted that the robot's speed was the most important safety feature (Report II, pp. 20, 24). Slower movement allows the robot to continuously scan its environment, observe its

surroundings, and avoid harming anyone nearby. In public spaces, controlling speed is therefore a simple but essential mechanism for maintaining safety.

A recurring insight in our empirical material is that robot safety is not determined solely by programming but is also highly dependent on the tools the robot is equipped with. While a robot may initially be classified as safe, this status can change if it is assigned tools that introduce new risks, such as knives or other sharp objects. In such cases, a new risk assessment must be conducted to account for the altered conditions.

This perspective was echoed by both Nick at ElectroFlight and employees at MakeBot (Report III, p. 9). At MakeBot, it was explicitly stated that robots themselves are not inherently dangerous; rather, it is the end-effectors or tools they operate with that ultimately determine their level of risk. Similarly, Nick described how safety procedures could become cumbersome when clients requested new functionalities, as each modification required a renewed risk assessment process, one that is both time-consuming and costly.

In addition to tool-related considerations, safety is also embedded in the robot's physical design. At MakeBot, for example, efforts were made to ensure a clean and simple shape to minimize potential hazards. As Martin noted, "the fewer edges, the safer it is" (Report III, p. 21), illustrating how even seemingly minor design choices can significantly impact safety.

Taken together, these examples highlight that safety is not confined to the robot's programmed actions but is equally a product of its material and structural design. Ensuring safety therefore involves a combination of software-based controls and deliberate physical design choices, both of which play a crucial role in mitigating risk.

Coordinating safety across systems: MES

A final example of formalised safety is found not in a single robot, but in the wider digital infrastructure that coordinates multiple robots, machines, and human interventions across the production environment found in some of our empirical cases. In Report III, this appears especially clearly in the *Manufacturing Execution System* (MES), which functions as a central coordinating layer in the production line. Rather than safety being located only in a robot's sensors, speed limits, or physical barriers, MES shows how safety can also be designed into the sequencing, routing, monitoring, and documentation of work across a larger socio-technical system.

MES tracks parts as they move through the line, receives updates from sensors and PLCs, and assigns tasks to different robots depending on the part's status and destination. In this way, it helps ensure that actions happen in the right order,

that parts are routed to the correct workstations, and that disruptions are managed before they cascade into larger problems. If one robot slows down, fails, or becomes unavailable, the system can redirect parts or adjust the pace of the line. The safety significance of this is not incidental. It lies in the fact that MES reduces risky misalignment between digital instructions, robotic actions, material flows, and human intervention by coordinating them in advance.

However, like many other technological systems, MES is not without limitations. During our fieldwork at MedCare, staff described how MES is used as a mediating layer between multiple distributed technologies. Despite this coordinating role, MES does not maintain full awareness of all system states. For instance, employees reported several situations in which an internal error within a robot caused it to stop functioning, without MES detecting the issue. As a result, MES continued to assign tasks to the robot, leading to the accumulation of a queue within the system. These problems typically only become apparent when an employee notices a robot positioned incorrectly and subsequently inspects its logs to identify the cause. In such cases, it becomes evident that while MES can register and display issues, it lacks the capacity to actively intervene or resolve them, and must here rely on humans.

MES also illustrates how formalised safety depends on traceability. Because the system records which robot handled which part, when a process occurred, and where problems emerged, it creates a documentary layer that supports accountability after the fact. In this respect, it resembles the logic already seen in risk assessment and standards: safe operation is not only a matter of preventing accidents in the moment, but also of making actions visible, recordable, and reviewable if something goes wrong. Here again, formal safety is inseparable from documentation.

What makes MES especially important for this chapter is that it shows how formalised safety in robotics increasingly operates at the level of distributed coordination rather than only at the level of individual machines. Safety emerges from the alignment of digital representations, system logic, physical movement, and human activity. A breakdown may therefore occur not only through the failure of a single robot, but through misalignment between several coordinated components. In this sense, MES exemplifies the most ambitious form of formalised safety: one that seeks to orchestrate complex interactions across a larger environment before risk becomes visible in practice.

At the same time, this section also marks the limit of formalisation and points directly toward the next section. Even a system explicitly designed to coordinate robots, parts, and human interventions safely cannot fully determine what happens in everyday work. Workers still encounter situations that require adjustment, improvisation, and judgment, and they still develop local ways of working with or around the formal system. MES therefore forms an ideal transition from safety as formalised to safety as enacted: it represents one of the strongest attempts to build safety into the system in advance, yet it also makes visible why safety in practice cannot be reduced to system design alone.

3. Safety as enacted

If the previous section showed how safety is formalised through standards, risk assessment, design, and system coordination, this section turns to the question of how safety is actually enacted once robots enter the contingencies of everyday practice. This is the ethnographic core of the chapter. Here, safety no longer appears primarily as a predefined property embedded in the robot or guaranteed by prior compliance. Instead, it appears as a situated

accomplishment produced through relations among humans, machines, infrastructures, routines, and judgments. Even highly formalised systems such as MES, which are explicitly designed to coordinate robots, parts, and human interventions safely, do not eliminate the need for local interpretation and practical adjustment. On the contrary, they make visible how much safe operation still depends on what workers, users, managers, and bystanders actually do in practice.

Themes in the empirical data

Across the cases, safety emerged as a shared analytical category because it repeatedly appeared as a site of learning. Not just for machines, but for humans and organisations learning how to live with robots in practice. What begins, in design and standards, as a technical requirement ("the robot must not hit anyone") becomes, in everyday use, a continuous cultural–material learning process in which people learn when to trust safety functions, when to override them and how. The vignette at the start of the chapter illustrates this complexity, showing that safety is not a preconditioned or deterministic property that can simply be imposed on robotic systems. Instead, it is something learned, expected and enacted through situated interaction between bodies, software, organisational routines, environments, regulations and standards. Across service, industrial, agricultural, and maritime settings, people learn and negotiate safety in action by engaging with breakdowns, anomalies, and workarounds—stretching rules, adjusting routines, and responding to moments where formal safety logic meets the unpredictability of real life. As preceding learning matters (Hasse 2020) many aspects can have a say in how people learn to adapt to new situations, including gender (Perez 2019). Seen through Hasse's lens of learning as cultural–material practice (2015), safety becomes a key site where autonomy is negotiated: where responsibility is relocated rather than removed, where human judgement remains indispensable, and where autonomy is revealed not as independence from humans, but as an ongoing, collective learning process embedded in socio-technical relations.

So safety is one of the main threads that ties all our cases together. Once you look across them, it becomes clear that "safety" is not just about avoiding physical harm, but about how robots, humans and organisational systems share risk and responsibility.

In the service-robot and cobot cases (restaurants, warehouses, conference robots), safety is initially framed very simply: the robot must not hit anyone. This leads to familiar design moves: speed limits, emergency stop buttons, light curtains, predefined routes and "safety bubbles" around the robot. In practice, however, the fieldwork shows that these measures quickly entangle with productivity and everyday work.

AGVs that are "too safe" (stop constantly, drive very slowly) become obstacles and are worked around, nudged, or ignored. Staff learn where they can "stretch" rules – to walk in front of the robot, to push it a bit, to override stops – and these micro-decisions become part of the real safety regime. What looks, in design documents, like a neat safety function becomes, in daily life, a shifting compromise between formal safety logic and informal work practices.

Across our cases, safety emerged as a site of learning. Machines learn and the humans and organisations learn how to live and work with machines under real conditions. What begins, in standards and design, as the relatively simple requirement that "the robot must not hit anyone" becomes, in use, a much more complex cultural-material process.

People learn where a robot has blind spots, how fast it stops, which signals they notice and which they stop registering, when formal safety measures help workflow and when they obstruct it, and when intervention becomes necessary. In this sense, safety is not merely implemented once and then left in place. It is continuously enacted through situated interaction between bodies, software, infrastructures, environments, and organizational routines.

This shift from specified safety to enacted safety is visible across all of the domains studied. In restaurants, museums, and retail environments, safe operation depends not only on sensors and slow speeds, but on how staff warn customers, reset robots, interpret delays, and manage congestion around them. In warehouses and industrial settings, safety is shaped not only by barriers, traffic rules, or interfaces, but by how workers adapt their movements to robotic systems and how they balance safety against pressure for speed, reliability, and throughput. In agriculture and maritime settings, safety is even less reducible to fixed rules or preprogrammed responses. Here, changing weather, uncertain data, animals, other vessels, and incomplete detection make human assessment, oversight, and practical judgment indispensable. Across these settings, safety does not disappear as a formal concern; rather, it becomes something that must be continuously completed in practice.

Seen through this lens, safety in the wild is not simply about avoiding physical harm in the narrow sense. It is also about how risk and responsibility are distributed across larger socio-technical systems. The empirical material shows that safety functions are rarely self-sufficient. A stop button only matters if people know when and how to use it. A traffic rule only stabilises movement if workers recognise it as workable. A fence or restricted zone only protects if it is respected, interpreted, and fitted into the rhythms of work. A coordination system such as MES only reduces risk if its alignments hold across the movements of robots, software, materials, and people. This is why safety should be understood as enacted: not because the formal layer is unimportant, but because it remains incomplete until it is taken up, lived with, and sometimes adjusted in concrete situations.

Trade-offs

During a robot conference, we spoke with a developer of safety sensors who explained that humans often bypass safety measures because they perceive them as slowing down work or reducing efficiency. This is a well-known phenomenon referred to as a “trade-off.” Trade-offs are a common concept in fields such as software and technology development, where designers frequently have to compromise certain functions in favour of others. These compromises can relate to usability, efficiency, flexibility and safety. One of the engineers in the project highlighted that a key contribution of Social Science and Humanities research is identifying these trade-offs. The example from the robot conference illustrates precisely this type of compromise in practice.

Philosopher Eric Hollnagel (2009) has explored trade-offs in relation to safety through what he calls the ETTO principle, which stands for Efficiency-Thoroughness Trade-Off. Hollnagel describes it as: “If demands for productivity or performance are high, thoroughness is reduced until the productivity goals are met. If demands for safety are high, efficiency is reduced until the safety goals are met.” (Hollnagel 2009, p. 15). According to Hollnagel, trade-offs are not unique to safety, but they occur across many aspects of daily life, as people constantly balance efficiency against

thoroughness. The decisions made in these situations are shaped by social norms, habits, and established practices. Therefore, trade-offs in robotics must be understood within the context of humans' daily operations and the difference between work as done versus formal procedures.

In our fieldwork, the concept of trade-offs emerged frequently, particularly in relation to safety. Many informants emphasized that compromising safety is never ideal, yet we observed it happening in practice. Humans often negotiate safety boundaries to adapt robots to their everyday routines. These compromises, whether conscious or unconscious, allowed for greater flexibility or efficiency but sometimes created situations where safety was not fully maintained.

Trade-offs also appear in the relationship between regulations, standards, and real-world practices. Compliance with safety standards is essential for selling robots, but many manufacturers, engineers, and customers perceive these standards as rigid and restrictive. As Nick from ElectroFlight noted: "It puts a stop block on a lot of things. Because we have so many different ways of doing various safety measures. But it's always connected with some kind of expensive cost or something that hinders production. That's how it is" (Report III, p. 9). Meeting these standards is time-consuming and expensive, and purchasing a robot involves more than just hardware costs. Even when human safety is prioritized, trade-offs occur in practice, highlighting the messy, imperfect social realities in which robots operate.

A particularly relevant trade-off is between safety and usability. Most robots come with a teach pendant, a handheld device used to operate the robot. At ElectroFlight, customers found the teach pendant difficult to navigate, so the company created a custom human-machine interface (HMI) that hides the pendant and simplifies operations (Report III, p. 5). While the teach pendant is designed to enforce safety, if it is too unintuitive, users may bypass it entirely. The custom HMI allows non-technical staff to operate the robot safely, providing flexibility while maintaining safety.

At MakeBot, engineers described a trade-off between sensitivity and reliability (Report III, p. 21). High sensitivity means the robot will stop at the slightest trigger, increasing safety but reducing reliability because it cannot consistently complete tasks. Lower sensitivity improves reliability but may reduce safety margins. This demonstrates that trade-offs can manifest in multiple dimensions of HRI.

MedCare faces a similar trade-off between safety and productivity (Report III, p. 49). Robots' maximum grip pressure is calibrated to prevent harm to humans, but this often exceeds the tolerance of the products being handled, resulting in damaged items. Adjusting the robot to handle products gently would make it overly sensitive and prone to stopping frequently during collaboration, illustrating the difficult balance between protecting humans and maintaining workflow.

At HandsON, the trade-off between speed and safety was also highlighted (Report III, p. 38). The AGVs could move faster, improving efficiency, but this would compromise employee safety. Mike suggested that this tension might explain why some employees occasionally defy traffic rules: impatience with speed limitations may push them to bend the rules while still feeling confident they can avoid harm.

Overall, these examples show that trade-offs are a fundamental aspect of HRI. Safety is never absolute; it is continuously negotiated alongside efficiency, usability, and reliability, reflecting the complex realities of robots operating in social and work contexts.

Adjusting to the Robot (Techno Regulation)

In many of the observed cases, ensuring safety did not rely solely on the robot itself, but required changes to the surrounding environment. At the restaurants, for example, small white markers were installed on the ceiling to help the robots determine their position and navigate safely. In this sense, safety is distributed across the entire setting, and the environment must be designed in ways that enable the robot to operate safely.

At ElectroFlight, this became even more explicit (Report III, p. 9). Nick described how safety measures such as infrared beams are used to detect movement and prevent accidents, emphasizing that: “Fences are not put up so that the robot cannot get out. They are put up so that employees cannot get in” (Report III, p. 9). While these barriers are clearly intended to protect humans from harm, they also function as physical constraints on human movement. In practice, safety is achieved by limiting access to the robot. Whether this is understood as restricting the robot or the human depends largely on perspective.

This approach echoes broader ideas about regulating technology through design. In their 2015 article, Leenes and Lucivero reference Isaac Asimov’s famous “Three Laws of Robotics,” which establish a hierarchy where robots must not harm humans, must obey them, and must protect themselves only insofar as this does not conflict with the first two laws (Leenes & Lucivero 2015). These principles place humans firmly at the top of the safety hierarchy and suggest that safety should be ensured by controlling robots to prevent harm. In practice, however, this often translates into controlling not only the robot, but also the human environment around it.

This form of regulation, often referred to as techno-regulation, operates by shaping or constraining human behaviour through technological means. As Leenes and Lucivero point out, this is not unique to robotics. Everyday examples include speed bumps, which are designed to enforce traffic laws by physically limiting how fast drivers can go (Leenes & Lucivero., 2015). In robotics, safety fences, sensors, and restricted zones similarly guide and limit how people move and act in relation to machines.

This differs from formalised safety measures, as they are constructed within social and negotiated environments. While elements such as safety fences constitute formalised equipment, we also examine other ways in which humans regulate their surroundings through technology. With the concept of techno-regulation in mind, our aim is not to analyse how these systems function in purely technical terms, but rather to understand the human and social motivations underlying

them. In this sense, the forms of techno-regulation discussed in this section differ from formalised safety measures, as they also emerge through situated practices rather than only predefined standards

A key question in this context is how much agency should be attributed to these technologies. While they appear to regulate human behaviour, they do not do so independently, but they are designed and implemented by people. This point is central to Langdon Winner's argument in "Do Artifacts Have Politics?" (Winner 1980), where he explores how technologies can embody social and political intentions. His example of the low-hanging bridges on Long Island illustrates how infrastructure can be used to exclude certain groups, in that case, preventing buses (and therefore primarily lower-income individuals) from accessing specific areas. Although the scale and context differ, the underlying principle is similar: technologies can shape, restrict, and organize human behaviour in ways that are not neutral.

In the case of robotic safety, fences and barriers are implemented to prevent harm, but they also require workers to constantly be aware of their movements and positions within the workspace. Safety, therefore, is not only about protecting humans from robots, but also about regulating how humans act around them. This highlights the importance of considering the broader consequences of such designs. Technologies are never neutral; they are created by people, for people, and inevitably reflect particular assumptions and priorities.

At the same time, developments in robotics have begun to shift these dynamics. As Rosa from MakeBot explained, earlier approaches to safety relied almost entirely on separating humans and robots (Report III, p. 23). Today, new safety systems allow for closer collaboration. While this might suggest increased human freedom, the examples above show that significant adjustments to the environment and to human behaviour are still required.

This is also evident in the warehouse setting at HandsON, where employees are expected to follow specific "traffic rules" designed to coordinate movement with the robots (Report III, p. 38). Although these rules might appear restrictive, employees described them as enabling rather than limiting. By clearly indicating where and when it is safe to move, the

rules create a sense of predictability and safety. Interestingly, this sense of safety can even lead employees to occasionally break the rules, as their confidence in the system gives them a feeling of control and autonomy.

These examples illustrate that techno-regulation does not simply constrain behaviour; it can also enable it. In many cases, safety measures provide workers with the knowledge and structure needed to navigate complex environments, thereby fostering a sense of freedom rather than limiting it. However, this comes with a trade-off. The robots used in a manufactures context are designed according to technical and safety standards that, in practice, shape how work must be carried out. Because these systems are often not flexible or tailored to local practices, it is the employees who must adapt their routines to fit the technology, rather than the other way around, which we will elaborate in the chapter on Tinkering where we also explore how humans help robots.

If trade-offs show how safety is balanced in practice, and techno-regulation shows how safety is distributed into environments and routines. The next section on accountability shows that responsibility for safe operation is redistributed rather than removed.

Accountability

Human unpredictability presents a central challenge in robotics, complicating questions of accountability within robotic systems. This connects to our earlier speculation on the development of accountable, self-adaptive robots, highlighting how accountability fundamentally shapes the ways in which robots operate within HRI. Regulations and standards play a crucial role in addressing this challenge. They not only hold developers accountable but also contribute to defining how responsibility is distributed across human and technological actors. When an unsafe incident occurs, the first step is to determine whether the robot underwent the necessary risk assessments and complied with relevant standards and regulations. If it did not, the developer may be held responsible. However, if the developer can show that all required safety measures were implemented, the investigation must then consider other potential factors. This process is not about questioning the ethics or morality of the developers, but rather about recognizing that safety operates within a structured system that extends beyond a simple binary of “safe” or “unsafe.” In practice, safety is negotiated in real-world contexts and must be understood in relation to the specific situation. This principle of distributed responsibility is mirrored in other domains, such as maritime navigation. Ultimately, the captain is legally responsible, as the rule of good seamanship ensures that someone is accountable in every situation (Report IV, p.19). An experienced captain, Frank, expressed scepticism about new technologies, arguing that they often fail to simplify life for ship crews. He outright rejected the idea of autonomous ships, claiming they would never exist because no software developer would be willing to take responsibility if such a ship were involved in an accident. For Frank, removing the captain creates a gap in accountability that current legal and regulatory systems cannot fill.

Cross-case analysis from our fieldwork further highlights that increasing robot autonomy does not remove human responsibility; it only redistributes it. Accountability is layered and shared across organizations, designers, and operators. Robots themselves cannot bear responsibility, as they lack agency. In hospitals and healthcare settings, managers and technicians remain answerable even when a robot or an automated system makes operational decisions (Report III, p. 44). On farms, farmers are still responsible for damage, animal welfare, and safe operation, even as they transition from direct control to remote supervision (Report IV, p. 26). At sea, maritime law explicitly assigns responsibility to human captains and shipowners, and there is currently no legal category for a truly autonomous vessel (Report IV, p. 17).

When we asked Henry from FarmGO what the most typical anomalies that could occur with this type of robot could be, he responded, “I don’t know, because… we’re not actually building the robot. We only make the… object recognition software” (Report IV, p. 6). In this statement, he is effectively delegating responsibility for risk assessment to the manufacturers, acknowledging that the developers of specific subsystems are not accountable for the robot as a whole. This illustrates an awareness among the different parties involved in robotics that responsibility is distributed.

From this, it becomes apparent that standards and regulations create a kind of loop of responsibility. Software or hardware developers ensure their components meet the necessary certifications, then pass the responsibility to the manufacturers, who in turn obtain their own certifications, and finally, the responsibility is passed on to the end-users. The users receive a product that meets all formal safety requirements, but how they choose to use it, or whether they follow safety procedures, remains largely outside anyone’s control.

Within the philosophy of technology, Don Ihde famously introduced the concept of the “designer fallacy,” which highlights that the way a technological artefact is designed does not determine how it will ultimately be used (Ihde 2006). The fallacy lies in assuming that design is neutral and that the intentions embedded within it will be interpreted uniformly by all users. From this perspective, it becomes clear that there is often a discrepancy between how designers intend for an artefact to function and how it is actually used in practice. Recognising this gap underscores the importance of attending to the situated and contextual realities in which technologies, such as self-adaptive robots, are deployed.

This is also the example used in the vignette of this section: “What they then come up with afterwards, just like if you drive your car and say, well, we can’t be bothered to use the seat belt. You can’t blame the manufacturer for that. It was there, it worked, but you’re not using it.” (Report III, p. 7). Here Nick from ElectroFlight explains that while he can provide all the necessary safety measures, he cannot control how, and if, the users follow them. In making this distinction, he is not denying responsibility but rather highlighting the complex challenge of assigning blame and ensuring accountability in HRI.

Rosa uses this same type of example, by comparing safety measures to producing a car with or without seatbelts: if drivers refuse to use seatbelts, they are endangering themselves, but if the car comes without seatbelts, they are forced to put themselves and others at risk (Report III, p. 24). It is by this logic that the company MakeBot ensures their robots

meet all safety requirements and provide guidance and training, but they cannot compel users to operate the robots safely. As she explained, “We have done everything we can to enable safety of those end users”.

This approach underscores a central insight from our research: while regulations and standards provide an essential safety framework, the ultimate effectiveness of these measures depends on human practices. Developers and manufacturers can build safety into the technology, but human behaviour, judgment, and the negotiation of trade-offs remain crucial in determining how safely robots are actually used.

Designers may describe scenarios where the “human is out of the loop,” but our research shows that safety-critical domains consistently keep humans firmly within a loop of responsibility. Technical questions, such as whether a robot can operate independently, cannot be separated from institutional questions about who is answerable when it fails. Autonomy shifts the distribution of responsibility but does not remove it. Humans remain central to accountability, oversight, and the establishment of trust in all safety-critical HRI.

Perceived Safety

At a conference, we spoke with Andrew, who is involved in a research project focused on improving safety in agricultural robots (Report I, p. 18). He explained that their approach to safety did not begin with technical specifications, but rather with the concerns of the farmers themselves, and what made them feel safe. This highlights an important distinction: safety is not only about physical protection, but also about how safe people feel in a given situation. We refer to this as perceived safety. Perceived safety is the subjective feeling of being safe in a given situation, which does not always align with the actual, measurable level of risk. While physical safety can be assessed through technical standards, regulations, and risk analyses, perceived safety is shaped by factors such as prior experiences, expectations, and how a system behaves or communicates. As a result, a system can be objectively safe while still feeling unsafe to users, making perceived safety a crucial aspect to consider in the design and implementation of technologies.

Andrew’s approach places perceived safety at the centre, emphasizing that understanding users’ worries is essential to designing systems they can trust. This perspective became very tangible in our own experience at a restaurant, where we tested a service robot by stepping in front of it to see how it would respond. Although the robot did stop and did not hit us, it did not stop as quickly as we expected. Despite the absence of any real danger, the interaction left us with a sense of unease.

This distinction was further reinforced in a conversation with Lucy, who works at one the restaurants (Report II, p. 9). She explained that she feels uncomfortable when the robots are moving around while she is alone. Objectively, this situation might even be safer, since there are fewer people, fewer obstacles, and fewer unpredictable interactions, for the robot to navigate. Yet her discomfort persists, which shows that it is not the actual level of risk that shapes her reaction, but her perception of it. Perceived safety is therefore highly context-dependent and can shift depending on the situation, even when the objective level of safety remains unchanged.

A similar point was raised during a seminar with engineers involved in the project, where we were discussing the concept of trust and perceived safety in robotics. One of the engineers opened up a discussion by asking how many people in

the room take the elevator to their office every day. When someone said they did, he followed up by asking, “Okay. Do you know when the last safety certification happened?” His point was that even when systems provide clear safety information, people rarely pay attention to it.

This example resonated with another one present at the meeting’s experience. Their office is on the sixth floor, and they used to take the elevator daily without thinking much about it. However, they said that recently they became aware of some unusual noises and movements while riding it. Since then, they had avoided using it, even though they rationally know that the elevator is certainly safe, and that it has passed all necessary inspections. However, that knowledge does not outweigh the uneasy feeling they got when they were inside it. Their sense of safety is shaped less by formal certifications and more by immediate, bodily experience. This illustrates the gap between physical safety (the actual, measurable level of risk) and perceived safety (the subjective feeling of being safe).

In many of the cases we observed in our fieldwork, the large red emergency button was highlighted as a crucial safety feature, especially in the contexts described in Report II. Workers explained that if something went wrong, or if they needed to stop the robot immediately, they could simply press this button. The button clearly provided a sense of safety and autonomy, as it allowed employees to take full control of the robot’s operation without needing to navigate complicated technical menus or settings. While it is unclear whether robot manufacturers fully endorse this practice, it was evident that workers had developed this as an internal work practice. The emergency button gave them a reliable and immediate way to feel secure and act safely in their interactions with the robots.

Just as a system can be objectively safe yet make humans feel unsafe, it can also be objectively unsafe while giving people a false sense of safety. This tension emerged repeatedly in our fieldwork and was often linked to what could be described as a form of defiance toward formal safety measures (Report II, III). While this will be explored in more detail in the chapter on “Sabotage,” it is important to consider here because it affects how safety measures function in practice.

At a conference, we spoke with employees from a company that designs safety sensors for robots (Report I, p. 19). They explained that their work is not only about making robots inherently safer, but also about compensating for human unpredictability. People often bypass or ignore safety measures such as fences, which creates additional risks. The company designs sensors to reduce accidents caused by such behaviour. When asked whether humans might be more cautious around very large industrial robots compared to smaller ones, the employee responded bluntly: “If you put up a fence, people will crawl over it. And so the larger the fence, the bigger the danger the humans are in.” This illustrates an important assumption in designing safety: humans cannot always be expected to comply with rules, so safety systems must account for defiance.

This tension of designing robots to be safe while interacting with unpredictable and sometimes noncompliant humans, emerged repeatedly in our observations. At GreenTech, for example, yellow lines marked a safety boundary around a conveyor belt, and the hoist was fenced in (Report III, p. 30). Despite this, employees routinely stood beyond the line, handling products as they were poured onto the belt. While no one was injured, these behaviours suggest that repeated safe experiences can lead workers to develop a relaxed or "loose" relationship with formal safety guidelines.

A similar pattern was observed at HandsON, where strict traffic rules are implemented to ensure safe robot navigation in the warehouse (Report III, p. 38). Employees often ignore these rules when they believe they can do so without harm. In both cases, noncompliance is not driven by a desire to be reckless, but by prior experience that nothing adverse has occurred. These patterns highlight that human behaviour is a critical factor in safety, and that technical safeguards alone cannot fully control risk. These examples also illustrate how human perceptions of safety can take many forms, and how these perceptions can strongly influence, and even shape, the way safety measures are implemented in robotics and in everyday life.

Service robots in public spaces, such as restaurants, stores, and museums, operate under unique constraints (Report II, p. 3). Because many humans in these settings have no prior experience interacting with robots, they may walk unpredictably, behave inattentively, or create hazards without realizing it. Robots in these environments are deliberately designed to move slowly, allowing sensors to detect obstacles and adjust accordingly. Staff, however, reported occasional collisions and even when no harm occurred, repeated minor incidents affected the staff's perception of risk, creating a paradox: humans sometimes ignored safety guidelines because they perceived them as unnecessary based on prior experience (Report II, III, IV). This phenomenon illustrates how perceived safety can diverge from objective safety and demonstrates the importance of understanding human behaviour in real-world contexts.

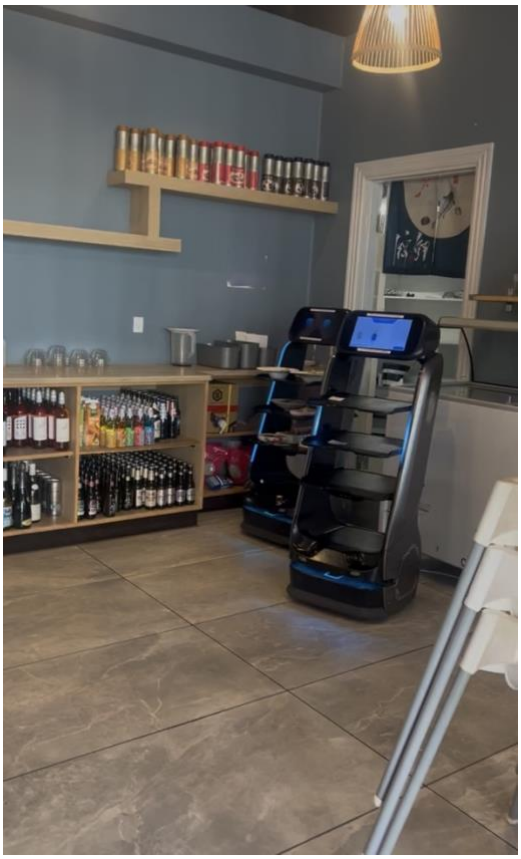
Safety not only as a formal feature – but negotiated in practice

As this report shows, formal safety procedures provide guidance and standards, but safety is always negotiated in practice. Human behaviour, context, and perception shape how rules are followed, and how safety is experienced in real-world interactions with robots.

In several service-robot cases, staff are responsible for managing safety in practice. This includes warning customers, "shepherding" robots through congested areas, and quietly resetting or moving them when they become stuck (Report II, III). These examples illustrate why, in this safety analysis, the gap between formal safety design, such as speed limits, stopping distances, and sensors, and actual safety practices is so important. Real safety depends on how staff and bystanders behave around robots and perceive hazards.

At the restaurants, staff noted that robots were not always fast enough to detect humans walking nearby, particularly when multiple robots were operating simultaneously (Report II, p. 15). This especially became a problem when they had

multiple robots going at the same time since the robots then had to coordinate and communicate with each other, and so they weren't always fast enough at detecting that someone had walked by. However, on the robot manufacturers website they emphasised the way in which the robots had sensors and coordination skills that would make them able to navigate safely among each other. This was just not the case in real time. So, although the robots were in fact able to coordinate with each other and slow-down in time, it became a problem when humans were in the mix and in social settings.



The same can be said about the little melody that the robot plays as it drives around. The robot plays this melody because it is otherwise very quiet, using the sound to alert nearby humans to its presence. However, Emma explains that one time she turned around and walked right into the robot, because she hadn't registered that it was standing behind her (Report II, p. 15). This shows that while the formal safety measure works as intended, in real social interactions it can fail, because Emma had become so accustomed to the melody that she didn't even notice it.

Other practical safety adjustments were also required. Robots were not allowed to carry used plates or drinks due to stability concerns, despite manufacturer claims that shock absorbers would minimize vibrations (Report II, p. 15). Staff and management actively negotiated safety in these situations, learning from experience what constituted a hazard in practice and adapting rules accordingly. These adjustments highlight that safety is not solely determined through risk assessments or formal standards: it requires continuous human observation, intervention, and judgment.

In the cases of FarmGO and ShipGO, the machines can never be fully autonomous (Report IV). This does not mean they cannot meet safety requirements, but human assessment and oversight remain essential, as real-world situations arise that must be managed by people. Safety is therefore a dynamic process, that is always partially up for negotiation, even when formal regulations and standards set the foundation.

This negotiation is not solely due to the complexity of real-world environments; it also reflects technological limitations. For example, Hanna at ShipGO explains that multiple technologies are necessary to reliably determine a ship's exact location (Report IV, p. 18). A single GPS alone could fail or misread, creating serious safety risks for such a large vehicle.

Additionally, concerns about hacking highlight that security is as critical as physical safety, reinforcing that achieving safe operation involves both technical reliability and human supervision.

Another challenge is ensuring that a ship avoids collisions with other vessels, illustrated by what the instructors at ShipGO call the "three-ship situation" (Report IV, p. 18) (Figure 2). Curt, one of the instructors, frequently poses this thought experiment to developers of anti-collision systems: it cannot be resolved with a single, deterministic rule, but instead must be addressed according to Rule two of the COLREGs (International Regulations for Preventing Collisions at Sea), which requires always demonstrating good seamanship and making correct judgments.

In this scenario, Ship A is on a collision course with Ship B. To comply with the rules, Ship A must turn to avoid Ship B, which may then put it on a collision course with Ship C. Ship C, in turn, must adjust its course to avoid Ship A, potentially moving into the path of Ship B, and so the cycle continues. Curt explains the practical implications: "So either you [Ship A] break the rules with regard to him [Ship B] and sail into him. Or you break the rules with regard to him [Ship C] when you turn towards him. Which, of course, is the right thing to do. Then the rules of the sea say that now he has a ship – now he has that white one there [Ship B] – a ship on his starboard side that he has to give way to. So he now has to turn over here" (Report IV, p. 18).

In this scenario, Ship 1 is on a collision course with Ship 2. To comply with the rules, Ship 1 must turn to avoid Ship 2, which may then put it on a collision course with Ship 3. Ship 3, in turn, must adjust its course to avoid Ship 1, potentially moving into the path of Ship 2, and so the cycle continues. Curt explains the practical implications: "So either you break the rules with regard to him and sail into him. Or you break the rules with regard to him when you turn towards him. Which, of course, is the right thing to do. Then the rules of the sea say that now he has a ship – now he has that white one there – a ship on his starboard side that he has to give way to. So he now has to turn over here" (Report IV, p. 18).

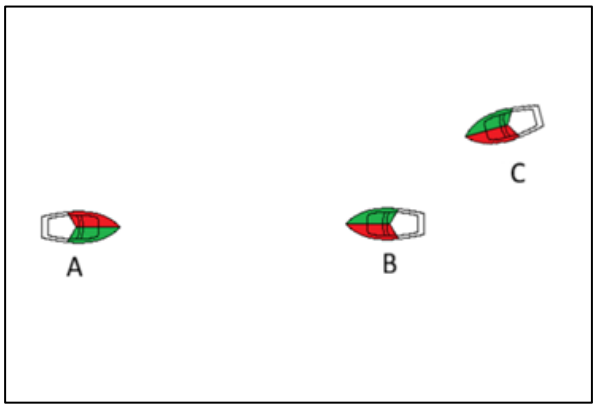


Figure 2: Three-Ship Situation

This example demonstrates that even in a highly regulated environment like maritime navigation, the rules cannot account for every possible scenario in the wild. Deterministic programming alone cannot solve the three-ship situation without violating some aspect of the rules, which is why the COLREGs emphasize exercising judgment and good seamanship. Such context-dependent decision-making is difficult to translate into robotic systems, highlighting how safety is not purely technical but negotiated and reliant on human judgment. In dynamic environments—whether at sea or in agricultural fields—the complexity increases, reinforcing that safe operation is a collaborative process between humans and machines (Report IV).

FarmGO's case also illustrate just how complex safety becomes when machines operate in dynamic, unpredictable environments (Report IV, p. 7). Their robots must distinguish between stones, weeds, humans, and fawns to avoid harm, yet they struggle to do so reliably. While a simple LiDAR sensor can detect that "something is there," this information is far from sufficient. The ethical stakes are high: a false detection could unnecessarily stop operations, while a missed detection could injure wildlife or damage expensive equipment. Safety here is no longer just a matter of "don't hit things"; it encompasses ethical, environmental, and material concerns (Report IV, p. 11).

Even if the system is designed for humans to be 'out of the loop,' human oversight remains critical. Operators are constantly needed to maintain the system's safety in practice, whether by intervening when the robot misidentifies objects or by resetting operations without overburdening themselves, safety is not just about preventing accidents, but about designing systems that humans can manage without constant interruption.

The challenge only multiplies in maritime contexts. ShipGO's navigation systems rely on multiple technologies because any single sensor, GPS, or chart can fail (Report IV, p. 18). In both agriculture and shipping, our fieldwork revealed that living beings, animals or humans, pose unpredictable safety challenges. At a conference, Andrew emphasized that farmers' primary safety concern was the robots' ability to identify fawns in the field (Report I, p. 18). Meeting this requirement meant improving sensor technology but also accounting for a level of situational awareness and understanding of animal behaviour that is extremely difficult to encode in machines.

Many real-world situations, such as the “three-ship situation,” cannot be managed simply by following rigid rules; instead, they require human judgment and sometimes rule-breaking in the interest of good seamanship (Report IV). Environmental factors like moisture, vibration, salt, waves, ice, and weather can degrade sensors and equipment in ways that are difficult to predict or model in advance. In these contexts, safety extends beyond mere collision avoidance, as it includes maintainability, the quality and reliability of data, and the capacity to make sound decisions when rules and instruments fail. This shows that safety is not purely a technical problem; it is inseparable from the context in which technology operates and the humans who interact with it.

What the fieldwork shows especially clearly is that enacted safety takes several recurring forms. First, it appears as trade-offs. Formal safety measures are constantly weighed against usability, efficiency, workflow, and reliability. A robot can be made highly sensitive and therefore highly interruptive; it can be made slower and therefore safer, but also more obstructive; it can be operated through interfaces designed for safety but experienced as difficult or impractical by users. In these situations, safety is not abandoned, but negotiated in relation to other demands. This is one of the clearest ways in which safety becomes a practical rather than merely technical matter.

Second, enacted safety appears through the adjustment of environments and human behaviour to robotic systems. In several of the cases, safe operation depended not only on the robot’s own design, but on reorganising the surrounding setting so that the robot could function safely within it. Markers on ceilings, safety zones, fences, restricted pathways, traffic rules, and regulated access all show that safety is distributed into the environment rather than residing only in the machine. Yet these arrangements do not simply constrain practice from above. They also become part of how workers learn to move, coordinate, and sometimes stretch the rules in relation to the robot. In this sense, safety is enacted through what the chapter terms techno-regulation: the shaping of conduct through designed infrastructures that both limit and enable action.

Third, safety remains inseparable from accountability.

One of the strongest findings across the cases is that increasing robot autonomy does not remove responsibility; it redistributes it. Developers, manufacturers, operators, managers, and end users all remain implicated in safe operation, though in different ways and at different moments. Formal systems may define responsibilities in advance, but accountability becomes especially visible when things go wrong, when systems behave unexpectedly, or when users act outside intended procedures. In those moments, safety is revealed not as a simple system property but as a field in which questions of responsibility, blame, and oversight are worked out in practice.

Fourth, enacted safety is shaped by perception. The fieldwork repeatedly shows that what matters is not only whether a situation is objectively safe according to standards and risk thresholds, but also whether it feels safe to those involved. Workers can know that a system is certified and still feel uneasy around it. They can also come to feel at ease in situations that are formally risky, precisely because nothing bad has happened before. Perceived safety is therefore not reducible to technical safety, but neither is it marginal to it. It affects trust, willingness to work with the system, attention to

warning signs, and readiness to intervene. It is one of the main ways in which safety becomes embodied and situational rather than purely formal.

Finally, and most broadly, safety is negotiated in practice. This is perhaps the central ethnographic finding of the section. Across the cases, workers and users do not simply follow formal safety procedures in a mechanical way. They interpret them, adapt them, supplement them, and sometimes bypass them. They learn what counts as "safe enough" in local conditions. They develop routines for resetting stuck robots, warning customers, steering machines through congestion, working around oversensitive stop functions, or balancing throughput against caution. Such practices are not exceptional deviations from a stable system. They are part of how safety is actually maintained in the wild. What appears in formal documents as a fixed safety function often reappears in practice as a compromise between formal logic and the demands of work as it is actually done.

Taken together, these dimensions show why safety in real settings cannot be reduced to a technical feature of robots alone. Safety is produced through situated relations among humans, machines, infrastructures, environments, and institutional arrangements. It is distributed rather than self-contained, and practical rather than merely specified. Rather than challenging the value of formal safety engineering, the ethnographic material shows what it must contend with in deployment: conflicting demands, imperfect environments, habituation, informal workarounds, bodily unease, distributed responsibility, and the practical judgments through which systems are made workable. The following subsections examine these enacted dimensions one by one, beginning with the trade-offs through which safe operation is continually balanced against efficiency, usability, and reliability.

4. Toward an empirical understanding of safety

Based on our empirical research into the understanding of safety, we argue that safety should not be regarded solely as a technical issue within the robot, something that can be resolved through formalised measures and techniques alone. Rather, we posit that safety emerges across three distinct but interconnected layers: first, as something imagined and promised; second, as something regulated and designed; and third, as something negotiated and enacted in practice. Taken together, these three layers require another understanding of safety. This empirical understanding does not reject engineering understandings of safety, nor does it reduce safety to subjective human experience alone. Instead, it shows that safety is produced through the ongoing relation between expectations, technical specifications, and situated practice. Safety is therefore best understood not as fixed, but as distributed, relational, and revisable.

The first layer, safety as imagined, shows that robots are encountered long before deployment through media, demonstrations, conferences, and public expectations. In these settings, safety is often presented as though it were an intrinsic feature of autonomous machines. Robots appear smooth, predictable, and naturally able to move among humans without friction. As the earlier section argued, this promised safety matters not because it explains how robots are actually kept safe, but because it shapes the horizon of expectations within which they are later encountered. It produces assumptions about what robots can do, how autonomous they are, and how safe HRI ought to feel. In this sense, imagined safety is not false, but incomplete: it frames the conditions under which disappointment, irritation, trust, or overconfidence later emerge.

The second layer, safety as regulated and designed, shows where engineering projects often begin. Here safety is translated into standards, risk assessments, validation and verification procedures, safeguards, interfaces, physical barriers, programmed limits, and wider systems of coordination. In this form, safety is something that can be documented, audited, and specified in advance. This is also the level at which RoboSAPIENS itself places much of its ambition, for instance through the Legitimize aspect of the MAPLE-K architecture, where planned actions are checked against relevant safety requirements. From this perspective, safety is indispensable as a formal achievement. Without this layer there would be no framework within which robots could be certified, trusted, or placed into use at all. Yet, as the chapter has also shown, even this formal layer does not eliminate interpretation. Risk assessments involve judgment, standards lag behind innovation, and systems such as MES, while highly sophisticated, still cannot determine practice completely in advance.

The third layer, safety as negotiated in practice, is where the ethnographic material becomes particularly informative. Across the cases, safety does not simply follow from prior design or formal compliance. It is continuously completed, adjusted, and sometimes stretched by the people who live and work with robots. Workers learn where robots have blind spots, how quickly they stop, when to trust an interface, when to use an emergency button, when to warn customers, when to reset a system, and when a formal safety feature becomes too obstructive to workflow. They also develop local understandings of what counts as “safe enough” under concrete conditions. In this sense, safety is not merely implemented and then left in place. It is enacted through trade-offs, practical adjustments, embodied judgments, distributed responsibility, and local workarounds. What appears in standards and design as a neat safety function often reappears in practice as a compromise between formal logic and the contingencies of work as it is actually done.

Seen together, these three layers suggest that safety is neither simply a matter of public expectation, nor fully a matter of formal design, nor only a matter of local practice. Safety is produced through the interaction between them. Imagined safety shapes what robots are expected to be. Regulated and designed safety translates those expectations into technical and legal forms. Negotiated safety shows how those forms are lived with, reinterpreted, and sometimes reworked in real settings. Through our empirical understanding, we posit that one must recognise that safety is partly representational, partly formal, and partly practical. Most importantly, it shows that safety is never finally settled, because the relation between these layers is constantly reopened as robots encounter new tasks, new settings, new users, and new failures.

This is also the point at which the chapter connects more directly to RoboSAPIENS. If RoboSAPIENS aims to develop self-adaptive robotic systems that remain trustworthy under unknown changes and variations, then safety cannot be understood only as one more layer of design. The chapter has already shown that safety must remain legible, interpretable, and revisable in deployment. That point becomes even more important in relation to self-adaptive systems. When systems change behaviour at runtime, the question is not only whether their actions remain formally within acceptable safety limits, but also whether those changes remain understandable to human actors, whether they can still be meaningfully intervened in, and whether organisations can learn from new frictions, breakdowns, and

unintended consequences. Perhaps, if one were to develop a revised notion of safety, it could be relevant to include runtime legibility, human oversight, and organisational learning as integral parts of safety itself, rather than as external additions to it.

From this perspective, safety is never final. The fieldwork repeatedly shows that new situations keep appearing: social settings become more crowded than expected, workers become habituated to warning signals, customers use systems differently than designers imagined, farms and ships confront changing environments, and technically safe systems become awkward, inefficient, or misleading in practice. These situations do not simply reveal that safety has failed. They also generate new routines, new workarounds, new local definitions of safe operation, and sometimes new demands for redesign. Safety is therefore revisable in a double sense: it is revised by practitioners in use, and it must be revisable by designers, organisations, and regulators in response to what deployment reveals. The significance of the chapter's findings lies precisely here. Safety is not only something specified before robots enter the world; it is something learned from what happens after they do.

From our empirical understanding, safety emerges as a socio-technical accomplishment produced across expectations, formal frameworks, and situated practices. It may therefore need to remain open to revision as these layers interact and are reconfigured under changing real-world conditions. Our empirical findings do not weaken the importance of standards, risk assessment, or technical safeguards. On the contrary, it makes their role clearer by showing that they are necessary, but not sufficient. It also clarifies the contribution of the ethnographic material to RoboSAPIENS: not to replace formal safety engineering, but to show the practical conditions under which safe autonomy can actually be sustained. In this sense, the chapter's main conclusion is that robot safety should not be understood as the elimination of uncertainty, but as the ongoing capacity to make systems safe, accountable, legible, and workable as new situations arise.

For RoboSAPIENS, this implies that future work on self-adaptive robotics should not only ask how systems can verify that an action remains formally safe, but also how they can support human understanding, intervention, and trust under changing conditions. Safety in this sense is not merely a boundary around adaptive behaviour. It is part of what makes adaptive behaviour usable and acceptable in the first place. A robot that adapts safely must therefore do more than avoid harm according to predefined thresholds. It must also remain intelligible to those who are responsible for working with it, capable of being interrupted or corrected when needed, and open to learning at the level of the organisation as well as the machine. This, finally, is what the three strands of the chapter together make visible: safety is not a static property of autonomous systems, but a revisable accomplishment of robots in the wild.

Discussion of Safety and Self-Adaptive Robotic Systems

Extending the discussion of robot safety into that of self-adaptive robotics and therefore the aim of Robosapiens, introduces a further layer of complexity, particularly when considering the possible influence upon human learning and the way trustworthiness verification is incorporated. Being able to continuously adjust their own behaviour in response to changing internal states and external environments, self-adaptive systems potentially shift safety, as it is technically

regulated and designed (ignoring here how it is negotiated by humans), from a largely pre-specified condition to an ongoing adaptive process.

If MAPLE-K systems capable of continuously sensing both the robot's internal state and its operational environment, are informing safety-relevant decisions, then this potentially raises challenges for human actors. Are they able to interpret not only the robot's actions, but also its adaptive logic? And equally for engineers and designers the question becomes, to what extent should the adaptive logic and monitoring data be made directly visible to users, so they can interpret how this data is being prioritised or understood. This becomes a marked extension to the complex issues of opacity or "black-boxing," where safety-critical processes become less accessible to users, fostering uncertainty around system predictability and reliability and thereby complicating trust.

The consequence of the robotic behaviour becoming less predictable can be that users either disengage (over-trust in the system) or conversely intervene excessively (under-trusting it), both of which have safety implications (the trust aspect of this will be unfolded in the next chapter on Trust). The knowledge base (the "K" in MAPE-K) is equally critical, as it potentially structures what the system "knows" about safe operation such as perhaps risk thresholds, and acceptable behaviours. This knowledge, however, is not static; it may be updated or reinterpreted over time, raising questions about how safety assumptions evolve and who is responsible for validating them.

For designers, of these self-adaptive systems there are further challenges. Traditional approaches to safety engineering rely on the ability to anticipate system states and certify behaviour against known conditions. Such as relevant standards and regulation procedures. Self-adaptive systems however disrupt this by introducing emergent behaviours that arise from runtime adaptation. Ensuring safety, therefore, requires new forms of verification that can operate dynamically rather than solely at design time. Designers must consider not only how to encode safety constraints, but how those constraints are maintained, revised, or even overridden as the system adapts.

Alternatively, because the MAPLE-K system can potentially incorporate safety enhancing loops of verification and trustworthiness checking into the adaptation cycle. This additional layer could have significant implications for robot safety. By continuously evaluating whether system behaviour remains within acceptable safety bounds - and by assessing the trustworthiness of its own adaptive processes -the system can, in principle, respond to anomalies, uncertainties, or degraded performance in more robust ways.

Continuous monitoring may also allow for the early detection of hazardous conditions in which adaptive planning can enable systems to respond to unforeseen risks in real time. Together with the ability for dynamic knowledge updating that allows safety models to evolve alongside changing environments. Seen in this light the MAPLE-K loop of verification and trustworthiness checking appears promising, as it is potentially capable of assessing the safety implications of its own adaptations.

However, realising such opportunities, requires careful attention to the human dimensions of negotiated safety highlighted throughout this section. Self-adaptive systems in themselves would not appear to eliminate the behaviour

of humans to engage in situated practices of negotiation, rather, it suggests these would be reconstructed and redistributed in new ways.

Conclusion on Safety

Taken together, this chapter shows that robot safety cannot be understood as a stable property located in the robot alone. Safety emerges across three interconnected layers: as something imagined and promised, as something regulated and designed, and as something negotiated in practice. It is therefore best understood not as fixed, but as a distributed and revisable socio-technical accomplishment.

The chapter has shown that formal safety remains indispensable. Standards, risk assessment, safeguards, verification procedures, and system coordination are necessary for anticipating harm, assigning responsibility, and making robots certifiable and usable. Yet the empirical material shows just as clearly that formal safety is never sufficient on its own. Risk assessments involve interpretation, standards lag behind innovation, and even highly coordinated systems cannot fully determine what happens once robots enter everyday settings.

In practice, safety is continuously completed by the people who live and work with robots. Workers, users, and managers learn when to trust a safety function, when to intervene, when rules can be stretched, and what counts as "safe enough" under local conditions. Safety is thus enacted through trade-offs, practical adjustments, distributed responsibility, bodily experience, and workarounds. What appears in standards and presented design as a fixed safety function often reappears in practice as a compromise between formal logic and the contingencies of work as it is actually done.

For RoboSAPIENS, the implication is that safety cannot be treated only as a design property to be verified in advance. If robotic systems are to adapt under unknown and changing conditions, safety must also remain legible, interpretable, and revisable in deployment. Safe autonomy therefore depends not only on technical compliance, but also on whether humans can understand system behaviour, intervene when needed, and learn from new frictions, failures, and emerging situations. Safety, in this sense, is not the elimination of uncertainty, but the ongoing capacity to make systems safe, accountable, and workable under changing real-world conditions.

This also points directly to the next section on trust. As the overall report argues, trust is closely tied to responsibility and accountability in practice, and the safety chapter has already shown why: whether people trust robotic systems depends not only on formal guarantees, but on how safe those systems feel, how understandable they are, and how responsibility is handled when things go wrong. Safety therefore prepares the ground for trust by showing the practical conditions under which confidence in robots can, or cannot, be sustained.

Trust

"[Context: Fieldwork at a navigation school, using different technologies to operate maritime vessels]

One of the problems with implementing new technology is that ship officers must both learn to trust the technology in order to use it, and at the same time not trust it too much, as it can also be wrong. One of the reasons for this is that when, for example, ECDIS was introduced, it was based on charts from 1949, whose data was based on measurements from before the 1920s. As Hanna described it: "The data was the same from 1970, and there was no better depth than what was measured back then. So there were actually, you could say, ECDIS-related groundings." This points to situations where people have had too much trust in a system that did not have better or more precise data than the charts they themselves had used for many years. Because they then possibly did not look out of the windows at their surroundings either but just assumed that ECDIS was a 1:1 version of reality, they then ran aground (Report IV: 19)."

[Introductory vignette drawing on Report IV illustrating several aspects of trust dealt with below.]

Introduction

Trust is where safety becomes experience, reliance, and accountability in practice. The previous chapter showed that safety in robotics cannot be reduced to formal guarantees alone, but must be continuously enacted, interpreted, and sustained in real-world settings. Trust begins from that point. It concerns not only whether a system meets standards or operates within acceptable limits, but whether people can and will rely on it in practice: whether it *feels* safe enough, whether its behaviour is intelligible, and whether responsibility remains clear when something goes wrong. In this sense, trust is not an addition to safety, but one of the ways safety is lived, judged, and made workable in everyday interaction.

The vignette above illustrates this clearly. Ship officers must learn to trust navigation technology enough to use it, yet not so much that they stop questioning it. In Hanna's example, over-trust in ECDIS could become dangerous precisely because the system's representation of the world was treated as more complete and reliable than it actually was. The issue was therefore not simply whether the technology worked, but how humans calibrated their reliance on it under real conditions. This points to a broader argument running through our cases: trust in robotics is never a stable yes-or-no attitude, but a situated and shifting judgment about when a system can be relied upon, when it must be checked, and when human intervention remains necessary.

This understanding resonates with the European Commission's Ethics Guidelines for Trustworthy AI. The guidelines argue that trustworthy AI must be lawful, ethical, and robust, and identify seven key requirements, including human agency and oversight, technical robustness and safety, transparency, and accountability. These requirements are useful here not because they offer a complete solution to trust, but because they show that trustworthiness is not reducible to technical performance alone. It also depends on whether humans retain meaningful oversight, whether systems are understandable in relation to their capabilities and limits, and whether their operation can be made accountable in practice. Our empirical material strongly supports this broader view. Across restaurants, warehouses, farms, hospitals, and maritime settings, people do not simply ask whether a robot functions. They ask whether it can be relied upon without doing harm, whether its decisions and data make sense, and who remains answerable when it fails.

In research on human–robot interaction (HRI), trust is frequently treated as a key determinant of how humans interpret, evaluate, and ultimately engage with robotic systems. Humans' trust in robotic systems is however a complex phenomenon that is shaped both by the ways in which these systems are designed, configured and reconfigured, and equally by the dynamic and unfolding nature of human engagement. The above extract encountered during our fieldwork illustrates how this complexity is implicitly embedded in everyday practice, and how trust in the context of HRI becomes a central factor in shaping human reasoning and action. More generally, it also fundamentally influences whether individuals are willing to engage with robotic systems at all.

More specifically the extract shows how in this case, the need to maintain a balance between too little and too much trust, highlights the inherently dynamic character of trust itself. In this sense, trust is not a fixed attribute of either the system design or of the human(s) involved, but is rather a situated phenomenon—emerging within particular contexts and evolving over time through social and cultural interaction. While the importance of trust in HRI is widely recognised, its dynamic nature and its configuration across different contexts require careful empirical and conceptual examination if more effective robotic systems are to be developed. This is particularly salient when considering self-adaptive robotic systems and the additional layers of uncertainty and reconfiguration that technologies such as MAPLE-K may introduce.

This chapter therefore approaches trust not as confidence in robots as autonomous beings, but as a situated calibration of whether humans can safely, intelligibly, and accountably rely on robotic systems in practice. Trust, as our cases show, is relational and situated rather than fixed; it is tied to responsibility and control, especially where humans remain accountable without fully controlling the system; it is shaped by technical opacity and uneven knowledge, so that both too little and too much trust become possible; and it is temporal, because it is built, broken, and recalibrated over time. Most importantly, trust is multi-layered. People may feel physically safe around a robot while distrusting its routes, representations, or decision-support.

In this chapter we explore this dynamic understanding of trust on the two axis; Trust about the robot, and trust in the robot. The first concerns the broader orientation people bring with them when encountering a robot, as well as how trust is formalised in design. This includes prior experiences with similar technologies, cultural imaginaries about robots and automation, expectations shaped by demonstrations or media, and confidence in the institutions, engineers, and regulations that stand behind the system. In this sense, trust about the robot is not produced in the situated interaction alone. It is already partly formed beforehand, through wider ideas of what robots are, what they are supposed to do, and what kind of technology they are taken to be.

By contrast, trust in the robot emerges in practice. It develops during concrete encounters in which people observe what the robot does, how it moves, how it responds, and whether it behaves in ways that can be relied upon in a given task and setting. This form of trust is therefore more immediate and situational. It is shaped not only by what the system is designed to do, but by how people interpret its behaviour under specific conditions and with specific risks in mind. Trust in this sense is never simply "there" as a fixed attribute of either the machine or the human. It is continuously calibrated through use, confirmed in some situations, weakened in others, and revised as people learn when the system can be relied upon and when it must be checked, corrected, or supplemented by human judgement. The distinction is

therefore not between an abstract and a real kind of trust, but between two temporal moments of the same process: one that precedes interaction and one that is enacted within it. Trust about the robot thus frames the encounter, while trust in the robot is what is negotiated through it. Together they help explain why trust in robotics is not reducible either to technical performance alone or to prior attitudes alone, but emerges through the relation between expectation and experience.

Seen from the perspective of RoboSAPIENS, this could be especially important. As the project aims to develop robotic systems capable of open-ended, trustworthy self-adaptation, then perhaps it is also relevant to consider the question of how adaptation remains interpretable, acceptable, and workable for those who live and work with it. Trust is therefore one of the central conditions under which adaptive autonomy can become usable in practice. It links directly back to the chapter on safety, since trust depends on whether safety remains legible and accountable in deployment, and it points forward to the following chapters on anomalies, Sim2Real, and Sabotage/Tinkering, where the practical limits of trust become especially visible. Trust, in short, is where adaptive autonomy is not only designed, but encountered, tested, and negotiated in the wild.

Trust about the robot

In this first axis, *trust about the robot*, it is important to begin with how trust is conceptualised within formalised and institutional frameworks. As introduced earlier, the EU's framework on Trustworthy AI provides a useful point of departure, as it explicitly addresses the relationship between humans and technology from a technical, ethical, and design-oriented perspective.

Building on this, we then turn to our empirical material to show how *trust about the robot* manifests in practice. Here, it takes the form of expectations, institutional confidence, and broader assumptions about technology. Together, these dimensions capture the temporal aspect of trust as something that is already partly formed before direct interaction takes place. In other words, they reflect how people approach robotic systems with pre-existing orientations that shape, and are later tested by, their experiences in practice.

Formalised trust

In an effort to reconcile our empirical findings with technical approaches to building trustworthy technologies, we have examined organisations that explicitly employ the concept of "trustworthy" in their frameworks. In this context, we turn to the EU's ethical guidelines for trustworthy AI. In 2018, the European Commission outlined a vision for supporting AI development in a manner that ensures "ethical, secure and cutting-edge AI made in Europe" (European Commission, 2019, p. 4). To advance this goal, the Commission established a High-Level Expert Group on AI, which in 2019 published the Ethics Guidelines for Trustworthy AI. The purpose of these guidelines is to promote the development and deployment of AI systems that can be considered trustworthy by providing a structured framework for achieving this aim (European Commission 2019, p. 2).

According to the report, trustworthy AI is grounded in three components: (1) it should be lawful, (2) ethical, and (3) robust (European Commission, 2019, p. 2). Importantly, the report emphasizes that while each component is necessary, none is sufficient on its own to ensure trustworthy AI (European Commission 2019, p. 2). Rather, they must function in a harmonically and overlapping manner. The guidelines further acknowledge that tensions may arise in practice when attempting to balance these components, suggesting that it is ultimately the responsibility of society to navigate and align them. This perspective resonates with our own findings, which highlight the complex and dynamic nature of real-world implementations.

In the following section, we will outline these guidelines alongside the Commission's explanations of each principle. We will then analyze them in relation to our empirical data and previously discussed examples, thereby illustrating the tensions that emerge in practice and how they are negotiated within society.

Seven key requirements for trustworthy AI (the list does not have a hierarchical order) (European Commission 2019, p. 14):

Human Agency and Oversight: AI systems should empower human beings so as to maintain human autonomy

Technical Robustness and Safety: they should be safe, reliable and secure.

Privacy and Data Governance: there should be respect for privacy and data, also with regards to quality and integrity of data.

Transparency: they should be transparent and tracible

Diversity, Non-discrimination and Fairness: they should foster diversity and avoid unfairness and bias

Societal and Environmental Well-being: they should be beneficial for current and future generations of humans

Accountability: there should be responsibility and accountability for the systems and their outcomes

Relating the Empirical findings to the EU Ethical Guidelines on Trustworthy AI

Across our empirical cases from reports (Report I-IV), we will now see how/if these dynamics shaping trust in robots can be interpreted through these requirements.

Human agency and oversight were particularly visible in cases where robots shifted existing power structures or redistributed responsibility. For example, in the cases of FarmGO and ShipGO, autonomous systems required humans, farmers and ship captains, to relinquish some level of direct control while still remaining responsible for outcomes (Report IV). This created tensions, as individuals had to trust the system while maintaining ultimate accountability. Similarly, at the shopfloor of MEASSURE, where a robot could potentially identify theft in a store, questions arose about whether management would trust the robot's judgment over that of human employees (Report II). These examples illustrate how trust depends on maintaining meaningful human oversight, even when systems operate autonomously.

The guideline concerning technical robustness and safety directly connects to the recurring concerns about physical safety observed in several field sites. At the restaurants and at MedCare, robots were designed to move slowly and stop when humans approached, signalling safety to users (Report III). However, informants also acknowledged that robots were not perfect and could have blind spots or malfunction. Stories of collisions or near accidents did not necessarily eliminate trust but instead reshaped how people interacted with the robots, reinforcing the idea that trust develops through experience with the system's safety mechanisms.

Transparency appeared repeatedly as a factor shaping trust, particularly in relation to the "black box" problem discussed by Nick from ElectroFlight (Report III, p. 11). Employees who did not understand how robots worked often found it difficult to trust them, especially when errors occurred. However, the cases also demonstrated that transparency is not simply about providing more information. As seen at MakeBot and FarmGO, too much information can lead to information overload or alarm fatigue, which may paradoxically decrease trust (Report III, IV). Transparency therefore requires not only openness but also communication that users can meaningfully interpret.

Accountability was also reflected across cases, responsibility for robotic actions was often shifted back onto humans rather than the technology itself. Customers in the restaurants blamed waiters rather than the robots when something went wrong, and Emma explicitly trusted the programmers behind the robot rather than the machine itself (Report II, p. 16). Similarly, Nick from ElectroFlight described trust in robots as essentially trust in the engineers who programmed them (Report III, p. 11). These examples highlight that trust in robotic systems is frequently mediated through trust in the human actors responsible for their design, operation, and maintenance.

The guidelines related to diversity, fairness, and societal well-being can also be connected to concerns about job replacement and changing work roles. At MedCare, fear that robots would replace workers had even led to sabotage of the system (Report III, p. 52). In contrast, other cases such as GreenTech and the restaurants showed more optimistic interpretations of automation as a tool that could assist workers rather than replace them (Report III, p. 31). These differing reactions illustrate how perceptions of fairness and social impact influence whether people trust or resist new technologies.

Taken together, the empirical cases illustrate that the EU principles for trustworthy AI are not merely abstract ethical ideals but correspond closely to the practical concerns people encounter when interacting with robotic systems. Trust in robots emerges not from a single factor but from the interaction between technical design, social expectations, responsibility structures, and human experience. As the fieldwork shows, building trustworthy robotic systems

therefore requires attention not only to technical robustness but also to the social, organisational, and cultural contexts in which the technologies are introduced.

Expectations, appearance, and disappointment

A further condition shaping trust across our cases is the relation between expectation and experience. Before people encounter a robot in practice, they often already have ideas about what it is, what it can do, and how autonomous or intelligent it is likely to be. These expectations are not formed in a vacuum. They are shaped by demonstrations, marketing, design choices, media imaginaries, and wider cultural narratives about robotics. Trust is therefore influenced not only by what a system does in practice, but also by the horizon of expectation against which that practice is judged. When robots are presented in ways that imply more competence, autonomy, or smoothness than they can actually sustain, disappointment can quickly weaken trust.

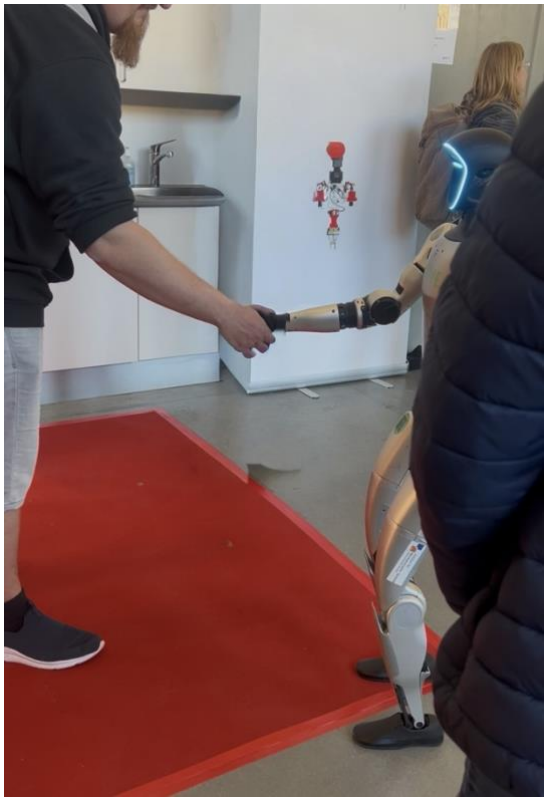
This dynamic became particularly clear during our fieldwork at a conference, where a two-legged robot attracted a large crowd by walking around, waving at spectators, and shaking hands. At first, the performance suggested a high degree of autonomy and sophistication. Yet the impression changed immediately when we noticed a man nearby holding a controller and closely observing the robot, apparently operating it remotely. What had seemed like autonomous action turned out to be staged and mediated by human control. The moment was analytically important because it showed how trust can be generated not only by actual capabilities, but by the presentation of capabilities. Once the presentation no longer matched the underlying reality, our expectations of the robot dropped accordingly. In this sense, disappointment did not arise from a technical malfunction, but from a mismatch between appearance and actual operation.

Nick from ElectroFlight described a similar problem from the other side. He explained that customers often expected the robots to be able to do far more than they actually could, which meant that early conversations frequently had to focus on managing expectations and clarifying limitations. His joking remark that the robot “can’t make coffee for you” is telling, because it points to a recurring practical task in robotics: not only designing systems, but also scaling back the imagined futures attached to them. Public demonstrations, trade fairs, and media representations may be successful in generating excitement, but they also risk producing unrealistic assumptions about autonomy, flexibility, and competence. Trust is then weakened not because the robot necessarily fails in an absolute sense, but because it fails relative to what people had been led to expect.

Appearance plays an important role in this process. Several of the robots encountered during fieldwork were clearly designed to invite trust through familiar or appealing forms. In the restaurants, robots had names and smiling digital faces, and Lucy explained that some restaurants even decorated them with bows or bowties. At MUSEA, the robot HUMANA was distinctly humanoid, with moving human-like eyes intended to create a sense of interaction with visitors. Such design choices are not superficial. They position the robot as approachable, friendly, and socially legible. But the material also shows that this strategy can work against itself. Anthropomorphic or human-like features may make robots seem easier to trust at first, yet they also raise expectations about how the system should behave. When the robot then fails to display human-like timing, knowledge, awareness, or movement, the mismatch can become more visible and

more disappointing. In these cases, human-like appearance does not stabilise trust so much as raise the threshold the system must meet.

The contrast with MakeBot is revealing here. Rather than aiming for human likeness, the company deliberately designed its robot to appear more toy-like while still remaining robust and functional. According to their experience, users were often more inclined to trust a robot that looked playful and non-threatening. The toy-like form appeared to lower expectations, making it easier for the robot to be accepted as a limited but useful technology rather than as an almost-human actor that ought to perform with human smoothness and competence. This suggests that trust is not always strengthened by designs that maximise social realism. In some cases, more modest and recognisable forms may be more effective precisely because they avoid overpromising what the robot can do.



The service-robot cases show that disappointment can also emerge from much smaller mismatches between expectation and experience. During our fieldwork in the restaurants, when we stepped in front of the robots to test their safety responses, the robots did stop, but not as quickly or as decisively as we had expected. Lucy likewise noted that she had initially assumed the robot could sense much more than it actually could. These are not dramatic failures, but they matter because trust is often calibrated through small, immediate encounters of this kind. Expectations need not be grand or futuristic in order to matter; even modest assumptions about sensing, timing, and responsiveness can shape whether a robot feels competent and trustworthy in everyday use.

At the same time, expectations do not operate only negatively. At MUSEA, HUMANA was explicitly used to create a “wow-effect” and to present the museum as technologically innovative. In that setting, the robot was not only an operational system but also a representative of the institution itself. Its appearance and presence were meant to generate confidence not just in the machine, but in the museum as a modern and forward-looking organisation. Yet this also means that the robot carries representational risk: if it malfunctions, misroutes, or collides with something, the disappointment may extend beyond trust in the robot alone and affect trust in the institution it is meant to embody. Expectations can therefore be productive, but only when the robot can sustain the role it has been given.

The agricultural and maritime cases show that expectations are shaped not only by design and spectacle, but also by work conditions and training. In agriculture, farmers are used to technological innovation, but they operate under significant economic pressures. If machines do not meet the expectations under which they were sold, this can directly affect productivity and income, which in turn weakens trust. Henry also noted that limited technical knowledge can encourage overly futuristic assumptions about what current technologies can realistically do. In maritime settings, the

pattern is almost the reverse. Hanna explained that operators are trained not to rely too heavily on certain measurement technologies because they are known to be imperfect. Here, expectations are deliberately lowered in order to prevent overreliance. Yet such caution can also generate scepticism from the outset. Across the cases, then, trust proves vulnerable to both kinds of mismatch: systems can be oversold and disappoint, but they can also be introduced so cautiously that people struggle to develop confidence in them at all.

Taken together, these cases show that trust is shaped not only by how robots function, but by how they are framed, staged, and imagined before they are used. Expectations are part of trust about the robot, but they are continuously tested in trust in practice. Where systems appear more autonomous, more human-like, or more capable than they actually are, disappointment can quickly weaken trust. Where expectations are managed more carefully, trust may be easier to stabilise because the system is judged against a more realistic understanding of its limits. Trust, in this sense, depends not simply on good performance, but on a workable fit between what the robot seems to promise and what it can actually deliver.

Trust in the robot

Trust in the robot concerns what happens when prior assumptions meet the realities of use, where opacity, bodily proximity, responsibility, breakdown, and practical judgement come into play. In these encounters, HRI is shaped by multiple, overlapping considerations: users' technical knowledge, questions of responsibility and control, and embodied as well as perceived senses of trust in the system.

These dimensions are not fixed to particular types of robots or singular experiences. Rather, as our fieldwork shows, they are dynamic and recurrent, emerging in different configurations depending on the specific context of interaction.

Opacity, technical knowledge, and transparency

A recurring theme across our cases is that trust is deeply shaped by what people know, what they do not know, and what they are able to understand about the systems they encounter. Mistrust often arose not because robots were visibly dangerous or obviously unreliable, but because their inner workings remained opaque to the people expected to live or work with them. Nick from ElectroFlight described this as a "black box" problem. When something went wrong, employees did not know how to interpret the malfunction, and unlike with a human colleague, they could not simply ask the robot what had happened. In his account, mistrust emerged not from hostility toward the technology as such, but from the difficulty of understanding and explaining its behaviour. At the same time, he also observed that trust could increase as workers became more familiar with the system and developed technical competence through use. Over time, employees learned more about what the robots could and could not do, and this growing competence strengthened both trust and collaboration (Report III, p. 12).

However, the cases also show that greater technical knowledge does not simply produce more trust. It can also produce overconfidence. Nick noted that engineers, precisely because they were so familiar with the robots, sometimes interacted too closely with them and became complacent about safety. What he called "habitual damage" points to an important asymmetry: too little knowledge can generate mistrust and avoidance, while too much familiarity can encourage a level of confidence that becomes risky in itself. Trust is therefore not a matter of obtaining as much

knowledge about the technological aspects as possible, but of developing a workable understanding of what the system is doing, what its limits are, and when reliance becomes over-reliance. This is one reason why trust in robotics repeatedly appears in our material not as a stable state, but as a calibration between too little and too much confidence.

The case of Emma in Restaurant 2 shows a related but slightly different dynamic. Emma did not claim deep technical insight into the robot's internal functioning. Instead, she trusted the people who had programmed it and therefore trusted that the robot would not do anything it had not been designed to do (Report II, p. 16). Here, opacity was not overcome by direct understanding of the system itself, but by confidence in the human actors behind it. Trust was effectively displaced from the machine to its designers and programmers. A similar pattern appears elsewhere in the chapter, where trust in robots often turns out to be trust in engineers, standards, maintenance, and institutional backing rather than in autonomous machine agency as such. Opacity, in other words, does not necessarily destroy trust, but it changes where trust is located.

At MakeBot, opacity took yet another form. Martin explained that his lack of technical knowledge made him reluctant even to touch the robot because he feared he might break it (Report III, p. 25). In this case, mistrust was directed not only toward the system, but toward his own ability to engage with it appropriately. The issue was therefore not simply whether the robot was trustworthy, but whether he himself felt competent enough to interact with it. This is important because it shows that trust in robotics is also tied to the user's sense of their own technological competence. Where systems appear too opaque or too fragile, hesitation can arise not from fear of harm but from uncertainty about how to act around them. By contrast, those with more technical knowledge may be willing to experiment more freely, which in turn can generate a more confident relation to the robot. Trust is thus shaped not only by what the robot is, but by the forms of literacy and confidence users are able to develop in relation to it.

This theme becomes especially clear in the agricultural material. Henry observed that farmers may be open to new technologies because they can see the practical benefits, but that this does not mean they are highly technologically skilled in a formal sense (Report IV, p.10). Because they manage many tasks simultaneously, systems must be intuitive and fit with already familiar interfaces and routines. This case helps show why transparency cannot be treated as a universal solution to mistrust. The issue is not merely whether more information is available, but whether that information is meaningful and usable in a specific practical context. A technically detailed system may be perfectly transparent in one sense, yet still remain effectively opaque to those who must use it under real working conditions.

For this reason, the empirical material repeatedly complicates the common assumption that transparency automatically builds trust. The chapter's earlier discussion of trustworthy AI already pointed in this direction: transparency matters, but only when it is tied to meaningful oversight and intelligibility in practice. In the trust material, this becomes clearer still. As the MakeBot and FarmGO cases suggest, too much information can produce information overload or alarm fatigue rather than confidence. Martin from MakeBot explained that some customers had asked for less information on the interface because the existing level of detail was too overwhelming. At FarmGO, operators had to monitor multiple screens with large amounts of data, making it necessary to prioritise which information actually mattered. When users are confronted with too many alarms, warnings, or notifications, the result may not be better understanding but

selective inattention. Transparency, in this sense, is not the same thing as dumping more information into the interface. It is a matter of making the right information available in a form that can be interpreted and acted upon.

However, technologies often function as "black boxes" for many users. While some are interested in opening these boxes to examine their inner workings, others are less concerned with technical details and instead seek to understand how technologies shape and affect the social world around them (Nye 2006). Engineers may assume that explanations or data displays will be meaningful because they are meaningful to them. Yet shared terminology does not necessarily imply shared understanding. As one conference encounter illustrated, an engineer found our arguments difficult to follow because we did not use the same English terminology as the technical community, while at the same time assuming that the concept of "trustworthiness" was self-evident across disciplinary contexts. This small episode is revealing because it shows that transparency is not simply a technical property but also a social and communicative one. What is visible to one group may remain obscure to another, even when both are looking at the same system. This highlights a recurring misunderstanding observed in our project: engineers tend to assume that shared terminology implies shared understanding.

What these cases show, taken together, is that trust is shaped not by transparency alone, but by interpretable transparency. People need not always understand every technical detail of a system in order to trust it. But they do need some workable sense of what the system is doing, what its limits are, and how they themselves are expected to relate to it. Where such understanding is absent, trust may weaken into caution, frustration, or avoidance. Where transparency overwhelms rather than clarifies, it may weaken trust just as much as opacity does. The issue is therefore not whether robotic systems should be made "transparent" in the abstract, but how their operation can be made intelligible for the particular people who must interact with them in practice. This is also why technical knowledge and transparency cannot be treated separately. What counts as a transparent system depends on who is using it, in what context, and with what responsibilities.

Responsibility, control, and professional identity

Across our cases, trust is closely tied to responsibility: not only to whether a robot works, but to who remains answerable when it does not. This question appears in two related forms. First, there is the question of accountability in a direct sense: when something goes wrong, who is held responsible for the robot's actions? Second, there is the question of control: can people trust a system if they remain accountable for outcomes while having only limited influence over how the system behaves? Taken together, these questions show that trust in robotics is never only a matter of trusting the machine itself. It is also a matter of trusting the wider arrangement of designers, operators, managers, standards, and institutions within which the machine operates.

This becomes visible already in the restaurant cases. At Restaurant 1, customers directed their frustration toward the human waiter rather than the robot whenever something went wrong with an order or service was delayed. In practice, the robot functioned mainly as a vessel carrying out tasks assigned by the staff, while the waiter remained the person held accountable. Even when the waiter had little insight into the cause of the problem, he still had to apologise and manage the consequences in real time. At Restaurant 2, Emma similarly apologised on behalf of the robot when it

bumped into objects or people, explaining that the fault did not lie with the robot as such, but with something malfunctioning beyond its control. In both cases, responsibility did not disappear because tasks had been delegated to a robot. It remained with the human worker, who had to absorb frustration, interpret malfunction, and preserve the relationship with customers. Trust in the system was therefore inseparable from the question of whether staff could meaningfully stand behind it in practice.

A related pattern appeared at MedCare, though here it was less the robot itself than the distributed structure around it that undermined trust. When errors occurred, responsibility was spread across a chain of actors, including technical support, distributors, and other organisational intermediaries. This made it difficult for employees to know who was actually responsible for identifying and resolving problems. In such situations, distributed responsibility did not strengthen trust but weakened it, because the system became harder to grasp and harder to hold together in practice. Yet the ShipGO case shows that distribution does not always have this effect. There, Hanna explained that they could not rely on a single GPS or navigation source, but instead depended on several technologies that could verify one another. In that setting, distributed infrastructure became a basis for trust because it allowed cross-checking and redundancy. The difference is revealing. Distribution fosters trust when it is intelligible and workable to those involved; it undermines trust when it disperses responsibility without making the system more understandable or controllable.

The ElectroFlight and HandsON cases sharpen this further by showing that responsibility and control can also come apart. Nick from ElectroFlight stressed that the company did everything it could to ensure that the robots were safe and compliant, but he also emphasised how difficult it remained to assign blame when things went wrong in practice. His comparison with a car and a seatbelt captures the point well: a manufacturer can provide safety mechanisms, but cannot fully control whether users will actually follow them. At HandsON, the opposite problem appeared. Workers trusted the robots' stopping mechanisms so much that they stopped respecting the traffic rules designed for safe collaboration. They stepped in front of the robots, expecting them to stop, and thus treated the system's safety features as a basis for bending the rules. In such cases, trust became destabilising rather than reassuring, because it encouraged people to rely on the robot precisely where the formal arrangement expected them not to. What these examples show is that trust is not simply a desirable attitude to be maximised. It must also remain aligned with the distribution of responsibility and control built into the system.

Responsibility also takes a second form in the material: the question of whether robots can be trusted to act in ways that match the standards, judgment, and values associated with human work. This is especially clear where robots begin to take on tasks that involve not just movement or delivery, but forms of decision-making, assessment, or delegated authority. At MEASSURE, for example, the robot was intended to identify theft by counting items and detecting discrepancies. Yet because the system was known to make mistakes, it introduced a new tension: if the robot indicated theft but employees disagreed, whose judgment would management trust? Here, the issue was not only whether the robot could technically perform the task, but whether its authority might begin to challenge or displace human judgment

in the workplace. Trust, in this sense, is not only about reliability. It is also about whether a system can be entrusted with forms of responsibility that have previously belonged to people.

This becomes even clearer in agriculture and maritime navigation, where professional identity plays a central role in how trust is formed. Henry from FarmGO explained that existing autonomous farming machines often still require farmers to walk alongside them "like a dog on a leash," and that one of the ambitions of automation is precisely to remove this need for constant supervision. But this shift is not merely technical. It also transforms the role of the farmer from direct actor to supervisor, and therefore requires trust that the machine can stand in, at least partly, for forms of judgment previously tied to professional responsibility. Andrew's example of farmers worrying whether robots can distinguish animals such as fawns from other obstacles makes this especially clear. Farmers do not only want a machine that works efficiently; they want one that can be trusted not to violate the practical and moral standards of good farming. Trust therefore depends not only on performance, but on whether automation appears compatible with what it means to be a competent and responsible practitioner.

The same is true, in an even sharper form, at sea. In the maritime cases, the captain remains the central legal and professional figure of responsibility. Good seamanship is not just a technical competence but part of professional identity itself. To delegate control to automated systems, captains must trust them deeply, yet the very weight of their responsibility makes that delegation difficult. As the material shows, captains are often sceptical not simply because they dislike technology, but because they must be certain that any system entrusted with navigation can meet the same standards of judgment and reliability that they are expected to uphold. Trust here is therefore not only about whether the ship can hold a course or avoid collision. It is also about whether there will still be someone to blame, explain, and repair when the system fails. This is why trust in autonomous ships becomes trust in a wider system of liability, maintenance, and oversight rather than in the vessel alone.

From this perspective, the theme of replacement belongs directly inside the question of responsibility and control. Fear of being replaced by robots is not merely a separate emotional reaction to new technology. It reflects a deeper concern that one's role, competence, and autonomy may be displaced while accountability remains. Nick from ElectroFlight observed that some companies hesitate to tell employees immediately when robots are being introduced because workers may read this as a threat to their jobs. At MedCare, such concerns had even led to sabotage, suggesting that mistrust was tied less to the robot's immediate safety or functionality than to anxieties about future redundancy. In contrast, staff in the restaurants did not believe waiter robots could replace them, since the machines could carry food but not perform the full range of service work. At GreenTech, the manager likewise saw robots less as replacements

than as tools for expanding production. These differences matter because they show that trust depends partly on whether automation is framed as deskilling and displacement, or as support and reconfiguration of work. Where robots appear to reduce people to "button pressers" or "robot babysitters," trust may erode because the system seems misaligned with what good work is supposed to be.

Taken together, these cases show that trust is inseparable from the relation between responsibility, control, and professional identity. People find it difficult to trust systems that leave them accountable without giving them meaningful control, that redistribute responsibility without making it intelligible, or that appear to undermine the standards and values through which they understand their own work. Trust in robotics is therefore not only trust in technical performance. It is also trust that the wider socio-technical arrangement remains answerable, workable, and compatible with the forms of judgment that matter in practice. This is why responsibility and control are so central to the trust chapter: they show that reliance on robots is always also reliance on the human and institutional structures that make those robots usable at all.

Physical and perceived trust in practice

A further layer of trust concerns the body. Across our cases, trust was not only articulated through what people said about robots, but through how they moved around them, how close they came to them, whether they hesitated, stepped aside, tested them, or ignored them. In this sense, trust is also perceived and embodied. It emerges through direct encounters in which humans interpret a robot's speed, proximity, movement trajectory, and responsiveness in relation to their own possible actions. Perceived trust is not a detached judgment but an interpretation of "the risk that something will happen" as interaction unfolds. Trust here is therefore closely tied to bodily experience: to whether a system feels manageable, whether it appears likely to stop in time, and whether its behaviour can be read well enough for people to orient themselves around it.

This bodily dimension is especially visible in the service-robot cases. In restaurants and similar public settings, first encounters with robots often involve curiosity, testing, or unease. Guests may approach the robot to see what it does, step in front of it, or keep a greater distance if they are uncertain. Staff, by contrast, often develop a more pragmatic relation to the machines over time. As the chapter's own synthetic formulation notes, physical trust reveals itself through body language: newcomers flinch or avoid, while experienced staff walk closer, cut in front of the robot, or lean over it because they have learned how it typically moves and stops. Trust here is not simply an inner attitude but a practical calibration enacted through bodily conduct.

The material on Lucy shows particularly clearly that physical trust does not map neatly onto whether someone has previously experienced harm. Lucy explained that her daughter had once been run over by one of the restaurant robots, and that Lucy herself had also been hit by the robots several times (Report II, p. 9). Yet these experiences did not lead her to avoid the machines altogether or treat them as fundamentally dangerous. Instead, she continued to work alongside them, sometimes even kicking or hitting them and jokingly referring to the robot as a "silent victim of violence." What this suggests is not that physical safety is irrelevant, but that trust in practice is graded and situational rather than binary. Even after unpleasant or risky encounters, people may continue to engage with the robot while

recalibrating their expectations and behaviour around it. Trust, in this bodily sense, is therefore neither all nor nothing; it is adjusted through repeated experience.

The design of movement plays an important role in this calibration. One of the simplest but most consequential ways in which the robots in Report II signalled safety was by moving slowly. In public environments, slow speed communicates that even if the machine is not fully understood, it is unlikely to cause serious harm by sudden impact. This is a case where the robot's behaviour itself becomes part of trust formation: people read speed as a sign of manageable risk. The robot may still be awkward, limited, or in need of supervision, but its slowness makes bodily coexistence more acceptable. Physical trust therefore depends not only on whether a system is formally safe, but on whether its movements can be interpreted as sufficiently predictable and non-threatening in practice.

The MedCare case shows the same point in a more complex workplace setting. Employees repeatedly demonstrated how the robots would stop when someone stepped in front of them and even encouraged others to test this for themselves. At the same time, they explained that the robots had a blind side that people should avoid, and told of a colleague who had once been run over and seriously hurt. Importantly, this story was not presented as proof that the robots were wholly untrustworthy, but as evidence that the systems were imperfect and therefore demanded caution. Employees thus learned to work with the robots by developing a bodily map of safe and unsafe positions around them. Trust was not based on an abstract belief that the system could never fail, but on an embodied understanding of where the limits of safe interaction lay.

A similar logic appears in the way employees talked about the more powerful robots in another restricted part of the warehouse. This area was off-limits because of the greater danger posed by the machines, yet the danger itself did not automatically produce rejection or mistrust. Rather, it seemed to make the area more intriguing and marked out a sharper distinction between robots that could be bodily shared with and robots that required separation. Physical trust, then, is not simply the absence of perceived danger. It is the practical sense that one knows how to move around the machine, how near to come, and under what conditions proximity remains acceptable. Where this practical sense exists, risk may be acknowledged without necessarily undermining trust.

The ElectroFlight case adds another important contrast. Nick emphasised that none of their employees had been harmed by the robot because they trusted, and followed, the safety mechanisms that had been put in place. In this case, physical trust is linked less to everyday improvisation and more to disciplined adherence to an already established safety regime. This is useful analytically because it shows that bodily trust can emerge in more than one way: either through situated learning around an imperfect system, as in the restaurant and MedCare cases, or through confidence in clearly structured safety measures that delimit bodily exposure from the outset. In both cases, however, trust is not just a technical property of the machine. It is a relation between robot movement, human perception, and the practical organisation of proximity.

The chapter's material on subjective experience sharpens this point further. During fieldwork at SHOPFLOOR, a customer walked past the MEASSURE robot and gave it the middle finger, even though the robot was moving slowly,

was surrounded by people, and gave no immediate sign of being threatening. When asked why, the customer simply replied that they “didn’t like robots.” This reaction is important because it shows that physical trust cannot be reduced to actual robot behaviour alone. Some forms of mistrust arise from prior dispositions, personal attitudes, or affective responses that persist even when the robot does not appear dangerous in any obvious sense. Perceived trust is therefore shaped both by the immediate behaviour of the system and by the histories, preferences, and sensibilities that people bring with them into the encounter.

Taken together, these cases show that physical and perceived trust in robotics is neither simply objective safety nor simply subjective feeling, but the relation between the two as it is lived in practice. People learn to trust robots bodily: by testing them, moving with them, noticing their rhythms, and adjusting to their limits. That is why physical trust often appears first in posture and conduct before it is formulated in words. A person who steps calmly in front of the robot, walks close beside it, or keeps distance from its blind side is already expressing a learned judgment about what kind of risk the machine poses. In this sense, physical trust is one of the clearest sites where safety becomes experience. It is where formal safety arrangements are translated into bodily orientation, and where trust becomes visible as a practical accomplishment of coexistence with robots in the wild.

Three layers of trust in practice

To further specify how trust in the robot unfolds in practice, this section draws together the empirical material into three analytically distinct, but interrelated, layers of trust. These layers do not introduce new categories external to the analysis; rather, they synthesise recurring patterns across the cases and clarify how practical trust becomes differentiated when people interact with robotic systems in concrete situations.

Across the material, trust in practice repeatedly crystallises around three questions. The first concerns physical trust: whether the robot can be approached without risk of harm. As shown in the public cases and MedCare, this form of trust is not established abstractly but through embodied experience, through how users interpret speed, proximity, stopping behaviour, and the robot’s responsiveness in shared space. Physical trust is therefore closely tied to perception and movement: it emerges as people learn how to position themselves around the system and calibrate their behaviour in relation to its limits. The second layer concerns functional trust: whether the robot can reliably perform the tasks it is introduced to carry out. Here, trust depends on whether the system proves useful within existing workflows, whether it reduces or introduces friction, and whether it can sustain performance over time. Breakdowns, inefficiencies, or mismatches between expected and actual capabilities do not necessarily eliminate trust, but they reshape it by forcing users to adjust how and when they rely on the system. The third layer concerns informational or representational trust: whether the data, models, and decision-making processes on which the robot operates can be relied upon. This includes trust in counts, maps, routes, and classifications, but also in the system’s broader representation of the environment. As several cases illustrate, even when a robot is physically safe and functionally useful, users may remain sceptical toward the validity of its outputs, especially when these are difficult to interpret or verify.

Importantly, these layers do not necessarily align. Trust may be established in one dimension while remaining fragile in another. A robot may be experienced as physically safe yet produce data that users do not fully trust; it may function

adequately within a workflow while still being perceived as opaque or unreliable in its decision-making. Trust in robotics is therefore not a singular or stable state, but a composite and shifting relation that should be continuously negotiated across different aspects of interaction. By distinguishing between physical, functional, and informational trust, the analysis makes visible how users evaluate robotic systems along multiple, sometimes conflicting, dimensions at once.

This layered perspective also provides a bridge to the broader question of self-adaptive systems. If trust in practice is already differentiated and situational under relatively stable conditions, then systems that change their behaviour over time may further complicate how these layers are established and maintained. The following discussion therefore considers how self-adaptation may reshape the conditions under which physical, functional, and informational trust can be formed, sustained, or challenged in practice.

Discussion of Trust in Self-Adaptive Robotic Systems

The preceding analysis has shown that trust in robotics is not a singular or static attribute, but a relational and situated accomplishment that develops through the interaction between humans, robots, and their environments. Trust emerges across expectations, embodied encounters, practical responsibilities, and ongoing interpretation. This is captured in the distinction between *trust about the robot*: the expectations and orientations that precede interaction and *trust in the robot:* the situated and continuously recalibrated trust that develops in practice. As the analysis has further shown, trust in practice tends to crystallise around three interrelated layers: physical, functional, and informational (or representational) trust.

This framework becomes particularly relevant when considering self-adaptive robotic systems. In the cases examined in this report, robots were generally encountered as systems whose behaviour, although sometimes opaque, was ultimately understood as relatively stable and grounded in predefined programming, safety mechanisms, and organisational arrangements. Trust was therefore often directed less toward the robot as an autonomous entity and more toward the engineers, standards, and institutions assumed to underpin it.

Self-adaptive systems, such as those envisioned through MAPLE-K, could complicate this relation. When a system is capable of modifying its behaviour over time in response to changing conditions, both predictability and responsibility become less straightforward to interpret. At the level of *trust about the robot*, adaptation may generate expectations that systems are capable of learning, improving, and reconfiguring themselves. While such expectations may strengthen confidence in the promise of adaptive autonomy, they may also introduce uncertainty regarding the system's limits and the locus of accountability.

At the level of *trust in practice*, self-adaptation introduces a further challenge: behaviour is no longer stable, but evolving. This can make it more difficult for users to establish the familiarity and patterns through which trust is typically calibrated. If a system changes its behaviour without clear explanation, users may struggle to determine when it can be relied upon and when intervention is necessary. Trust, in this sense, could become harder to stabilise because the conditions under which it is formed are themselves in flux.

These challenges extend across all three layers of trust identified in the analysis. In terms of physical trust, adaptation may affect whether users can anticipate how the robot will move, stop, or respond in dynamic environments. In terms of functional trust, it raises questions about whether adaptation enhances or undermines reliability in task performance. In terms of informational or representational trust, it concerns whether users can rely on the system's evolving data, models, and decisions, particularly when these are updated continuously rather than remaining fixed.

However, self-adaptation does not necessarily undermine trust. It may also provide new grounds for it. If adaptive processes are made intelligible and meaningful in practice, they can be interpreted as signs of competence rather than unpredictability. Systems that are able to detect anomalies, adjust to changing conditions, and improve their performance may strengthen trust, provided that users can understand and relate these changes to their own practical concerns.

The central issue, therefore, is not adaptation as such, but how adaptation is experienced. The empirical material throughout this chapter has shown that trust depends not only on technical capability, but on interpretability, usability, and accountability in context. A self-adaptive system that changes behaviour without explanation may be experienced as opaque and unreliable. Conversely, if adaptation is communicated in ways that align with familiar practices of learning, correction, and adjustment, it may become a resource for trust rather than a source of uncertainty.

From this perspective, the challenge of trustworthy self-adaptation cannot be addressed through technical robustness alone. It requires that adaptive behaviour remains legible, that users can recognise what the system is doing, and that responsibility remains interpretable when outcomes are uncertain or contested. Trust, in this sense, is one of the practical conditions under which adaptive autonomy becomes usable.

For RoboSAPIENS, this may suggest that the development of self-adaptive robotic systems cannot rely only on runtime verification and technical performance alone, but might also need to account more explicitly for the social and organisational contexts in which these systems are deployed. Adaptive systems do not remove the human dimensions of trust identified in this chapter; they intensify them. The question is therefore not simply whether robots can adapt, but whether their adaptation can be integrated into the practices through which trust is formed: practical experience, intelligibility, responsibility, and the ongoing calibration between too little and too much trust.

Understood in this way, self-adaptive robotics reopens the central question of this chapter: under what conditions can humans safely, intelligibly, and accountably rely on robotic systems in practice? The analysis suggests that adaptation becomes trustworthy not when uncertainty is eliminated, but when it remains interpretable and manageable within the socio-technical settings in which humans and robots coexist.

Conclusion on Trust

Taken together, this chapter shows that trust in robotics cannot be understood as a fixed property of either the robot or the human user. Across our cases, trust emerges as a relational and situated accomplishment that is continuously formed, tested, and revised in practice. People do not simply decide once and for all whether they trust a robot. Rather, they calibrate trust through prior expectations, lived encounters, practical responsibilities, and embodied experiences

of working and living with robotic systems. In this sense, trust is not an abstract attitude added on to technical performance. It is one of the practical conditions under which robotic systems become usable, acceptable, and workable in real-world settings.

The chapter has argued that this process can be understood through the distinction between *trust about the robot* and *trust in the robot*. *Trust about the robot* refers to formalised conceptions of trust, as well as the expectations and representations that shape how robots are perceived before interaction. *Trust in the robot*, by contrast, emerges within concrete situations, as people observe robotic behaviour, interpret its actions, assess risks, and adjust their own conduct accordingly. These two moments are closely intertwined and are primarily distinguished by their temporal orientation - one preceding interaction, the other unfolding within it. Trust in robotics thus develops in the space between what a robot appears to promise and what it proves capable of sustaining in practice.

The chapter has also shown that trust is shaped by a number of recurring conditions across layers. Opacity and uneven technical knowledge affect whether users can interpret the system meaningfully. Responsibility and control matter because people often remain accountable for outcomes without fully controlling system behaviour. Expectations matter because robots are frequently encountered through promises, designs, and imaginaries that may exceed what they can actually do. Bodily experience matters because trust is also learned through movement, proximity, and repeated encounters with risk. And temporality matters because trust is built, broken, and recalibrated over time. Trust, in this sense, turns out to be less about whether people simply “like” robots and more about how they learn to live with systems that are partly opaque, sometimes unreliable, and deeply entangled with their own professional identities, practical responsibilities, and everyday risks. For that reason, the chapter also develops trust through three interrelated layers: physical trust, functional trust, and informational or representational trust. This makes it possible to show more precisely how trust is formed, strained, and redistributed across different human–robot settings.

This conclusion also clarifies the chapter’s place within the larger argument of the report. The safety chapter showed that robotic systems must remain legible, accountable, and workable in deployment, and trust begins precisely there. Trust is what happens when those safety arrangements are encountered, interpreted, and lived with by workers, users, managers, and bystanders. In that sense, trust is where safety becomes reliance in practice. At the same time, the chapter points forward to the following sections. Trust is repeatedly tested where anomalies appear, where simulated models fail to align with lived environments, and where humans intervene, tinker, or resist. Trust therefore links directly to the practical limits of autonomy explored in the later chapters.

For RoboSAPIENS, this is relevant, as if the project seeks robotic systems capable of open-ended, trustworthy self-adaptation, then perhaps trust should not only be treated as a simple outcome of technical robustness alone. Trustworthy adaptation depends on whether systems remain intelligible, acceptable, and workable for the people who encounter them. This is especially important for self-adaptive systems. If robots can modify their behaviour over time, trust will depend not only on whether those adaptations remain formally safe, but on whether humans can understand them, judge them, and intervene when necessary. Trust in adaptive robotics must therefore be thought of not as

confidence in autonomous machines as such, but as an ongoing calibration of whether humans can safely, intelligibly, and accountably rely on them in practice.

This distinction is especially useful in relation to the empirical material. Across our cases, people often approached robots with expectations already shaped by appearance, prior experiences, institutional assurances, and wider narratives about automation. Yet these expectations were then adjusted, sometimes strengthened and sometimes unsettled, once the robot had to function in practice. A robot may initially be trusted because it appears friendly, because it is assumed to be carefully designed, or because it is associated with a reputable organisation. But such trust remains provisional until it is tested in everyday interaction. Conversely, people may begin from scepticism or uncertainty, yet gradually develop trust through repeated encounters in which the system proves understandable or workable

Anomalies

"*One evening during our fieldwork in a restaurant, a waiter dropped glasses on the floor and began sweeping the shards into a pile. A service robot approached without registering the hazard. The staff pressed the emergency stop button to prevent it from driving straight through the pile, then cleared a path and restarted the robot. One waiter kept a hand on the top tray—ready to stop it again, or to ensure it did not turn into the shards. Moments later, a second robot drove directly through the remaining pile, spreading shards further and forcing renewed intervention: stopping the robot, removing shards from under it, and sweeping in front of it so it could continue. In that moment, "autonomy" was not a property of the robot alone; it was a fragile achievement sustained through human vigilance, quick judgment, and coordinated repair work. (Report II, p.8)."*

[Introductory vignette drawing on Report II illustrating several aspects of anomalies dealt with below.]

Introduction

This is an episode from our fieldwork in which anomalies become central to human action. Although at first the above extract might be understood as a simple technical fault in which the robot is unable to recognise the common feature of broken glass often found in such environments. Rather than an error, such an instance is better understood as a kind of relational disruption between robot and human worlds. The episode therefore reveals the way in which deviations become analytically valuable to understanding limitations in human-robot interactions (HRI). In this example anomalies can be seen as moments when the practical limits of autonomy become visible, because the robot's expectations meet a world that can only partially be modelled and in which humans must step in to re-stabilise safety, workflow, and meaning. Anomalies are therefore a deviation between what is sensed and reasoned by the robot and by humans respectively. Or in other words deviations from expected system behaviour or interactional norms that become meaningful through the situated interpretation of human actors, and that may arise from misalignments between computational models, physical performance, and socio-cultural expectations within human–robot systems.

Across our cases, anomalies are where the real character of robot autonomy shows through. They are the moments when systems behave in ways that were not planned or expected and they can be both deeply problematic and unexpectedly productive. Looking across the material, anomalies are not rare glitches at the edge of otherwise smooth operation. Rather, they are central to how systems are understood, tested, distrusted, repaired and improved in everyday practices.

In the engineering literature, anomalies are often defined in data-driven terms. The taxonomy paper used in RoboSAPIENS defines an anomaly as "an observation or a set of observations within a dataset that markedly deviate from established norms or expected patterns," (Moustakidis, Passalis & Tefas, 2024, p 6.) and it stresses that both significance and interpretation are context-dependent; detection involves interpretation as well as identification. This

definition aligns with a technical view of autonomy as continuous monitoring of state and environment for deviations that might require correction or adaptation.

Our fieldwork however incorporates the human perspective and shows that anomalies do not simply exist in data streams. They also exist in practice: as situational breaks where something becomes unsafe, confusing, inefficient, or incoherent—sometimes even when the robot registers no deviation at all (as in the glass-shard vignette above). A key contribution of this chapter is therefore not to reject the technical definition, but to extend it: to show how technical anomaly detection and practical anomaly experience, form two different "registration systems" that do not reliably overlap.

We therefore construct the chapter into two perspectives; (1) from where the anomaly is perceived and registered, and (2) through their valence, meaning the effects they produce in situated practice. These two distinctions are analytically separate but constantly intertwined in real-world operations. The result is not a fixed taxonomy of anomaly types, but a relational field in which deviations only become meaningful through their location in both registers and through their practical consequences.

It is important to emphasise that this distinction is made from an empirical perspective, focusing on how anomalies are enacted and interpreted in practice, rather than from a technical account of anomaly detection as a computational process. When we refer to an anomaly being "detected from the robot's perspective," this should not be understood as a claim about the internal mechanics of detection systems. Instead, it refers to situations where the robot demonstrably reacts to something as if it were anomalous - through stopping, re-planning, alerting, or altering its behaviour - while this same event is not necessarily experienced or recognised as an anomaly from the human point of view. In other words, the distinction is grounded in observable differences in how humans and robots respectively register and respond to the same situation, rather than in any detailed specification of how anomaly detection is implemented within the system.

Valence

Across the empirical material, anomalies can be understood through their valence, meaning the effects they produce in situated practice. This distinction does not imply that anomalies are inherently positive or negative, but rather that their meaning emerges through how they are interpreted, managed, and integrated into ongoing work. The same deviation can therefore shift valence depending on context, timing, and the available capacity for response.

Valence, in this sense, is not a property of the anomaly itself, but a relational outcome of HRI. It expresses whether a deviation is experienced as a disruption that must be corrected, or as a resource that enables learning, repair, or adaptation. Importantly, this means that valence is not stable or fixed at the point of detection. Instead, it is continuously produced in practice, as actors respond to, negotiate with, and make sense of unfolding system behaviour.

In the following sections, we therefore distinguish between negative and positive anomalies. These categories should not be understood as objective or intrinsic states of the anomaly itself, but as analytical descriptions of how deviations are experienced and enacted in specific empirical situations. The distinction is thus not final or exhaustive, but a way of

tracing how the same type of system deviation can take on different meanings and consequences depending on the social and material context in which it appears.

Negative anomalies are those moments in which deviation produces immediate disruption of workflow, safety, or coordination, and therefore requires intervention to restore stability. They are characterised by a breakdown in alignment between system behaviour and practical expectations, where action must be redirected, paused, or corrected in order for work to continue. Across the cases, negative anomalies are defined less by technical malfunction in a strict sense than by their practical consequences: they interrupt coordination, introduce uncertainty, increase cognitive and physical workload, and often force reactive forms of human intervention. In many situations, they also expose the limits of system perception or decision-making, making visible what the system cannot adequately account for in real time.

Positive anomalies, by contrast, are those deviations that do not only disrupt but also become productive within practice. Rather than being resolved solely as problems, they open up possibilities for learning, repair, and adaptation. They may reveal hidden assumptions in system design, expose misalignments between models and reality, or generate opportunities for improving workflows and procedures. In some cases, they stabilise operations through informal adjustments and workarounds; in others, they contribute to longer-term redesign and system refinement. Positive anomalies therefore highlight that deviation is not only a source of breakdown, but also a condition for the emergence of new forms of knowledge and practical improvement.

This is compatible with arguments made in the article “Trust and anomalies in the design of safe robotic self-adaptation”, made by a group of engineers in RoboSAPIENS together with anthropologists and a philosopher. In this cross-disciplinary article we argue that that labelling something anomalous assumes a baseline of “normalcy,” and that in practice “normal” is co-produced in situ, which raise the question of who’s normal is embedded in design and monitoring. (Birch et al., 2026, p.5).

Taken together, the distinction between negative and positive anomalies underscores that deviation in human–robot systems is never purely technical. It is always interpreted, situated, and acted upon within specific material and organisational conditions, where its meaning and consequences are continuously negotiated.

Robot-perceived anomalies

From the perspective of robot perception, anomalies are first and foremost defined by system-internal logics such as sensor thresholds, probabilistic models, classification rules, and expected state transitions. Within this register, anomalies appear as deviations from learned or programmed normality. Importantly, however, what counts as an anomaly for the system does not necessarily correspond to what is experienced as meaningful disruption in the environment. The analytical point here is not to describe how anomaly detection works technically, but to focus on situations where the robot acts as if something is anomalous, through stopping, re-planning, alerting, or altering its behaviour, without this necessarily being recognised as such from the human perspective.

Negative anomalies

A large and recurring category of robot-perceived anomalies produces negative practical consequences in human work. This is most clearly visible in cases of false positives, where the system identifies an object or situation as hazardous or abnormal—such as misclassifying a stone as a fawn or straw as a person—and responds by stopping, issuing alerts, or interrupting ongoing processes. In these situations, the anomaly is operationally real within the system's internal logic, but it does not correspond to a meaningful deviation in the world as experienced by human actors. The result is not improved safety in any immediate sense, but rather disruption: unnecessary halts, additional verification work, and the reintroduction of human supervision into processes that are formally designed to be autonomous.

In practice, these events generate a form of friction in everyday workflows. Even when system responses are justified within a safety-oriented design logic, they can undermine efficiency and create uncertainty about when intervention is actually required. Over time, repeated false positives may shift the role of human actors from occasional overseers to continuous validators of system output. In this sense, robot-perceived negative anomalies do not simply represent "errors," but become structuring elements of labour redistribution between human and machine. A central example from FarmGO is deceptively simple: the system "thinks" it sees a fawn, but it is a stone; it detects straw as if it were a person. The machine reacts by stopping yet the stop may be unnecessary because the anomaly is a false positive. The farmer receives an alert on the screen and must go out to check. This is where the promise of autonomy becomes practically contested: if the whole idea is that the farmer should not have to walk with the machine, repeated false positives pull the farmer back into constant supervision and wasted trips. (Report IV, pp.4–5; pp.9–10).

The researcher in farming systems, Henry, makes the economic and practical point explicit: if the robot is not efficient, it becomes a poor investment, and the idea of distancing the farmer from the machine "disappears," because the farmer must continually go out to see why it stopped. (Report IV, pp.9–10). Analytically, this case is a strong example of how anomaly management is never just technical: it directly reorganises time, labour, and the credibility of autonomy. It also illustrates the engineering taxonomy's "normal state with anomalous observation" (Moustakidis, Passalis & Tefas, 2024, p.10): the observation pipeline produces a stop-worthy signal even when the world-state does not warrant it, (Report IV, pp.4–5).

A closely related form involves unnecessary system interruptions triggered by ambiguous or borderline inputs, where the system defaults to caution in the absence of clear classification. While such behaviour may be designed to maximise safety margins, it often results in operational over-sensitivity. From a practical standpoint, this means that autonomy is partially offset by recurrent stoppages that require human confirmation before work can continue. Rather than reducing human involvement, the system reintroduces it in fragmented and reactive forms.

At HandsON, one AGV stopped 1–2 meters before a conveyor and stood still for several minutes. The stoppage became consequential only when more and more AGVs lined up behind it, forming a queue. The fix was simple: the responsible employee pressed a button, pulled the AGV about one meter backwards, and restarted it; the queue dissolved. Fiona summarised the everydayness of such events: "We're surprised when they work." (Report III, pp.35–36).

The point is not that engineers "missed" the problem, but that repeated stoppages can become operationally normalised, treated by employees as background noise, especially when humans learn to resolve them quickly. This supports the taxonomy's broader claim: separating normal variability from "true anomalies" is itself interpretive and context dependent (Moustakidis, Passalis & Tefas, 2024, p.7).

Positive anomalies

Robot-perceived anomalies can also have positive practical consequences, even when they originate as deviations from expected system behaviour. In these cases, the value of the anomaly is not that it is "correct" in a technical sense, but that it becomes productive within human interpretive and organisational practices.

One important form concerns early warning effects. A system stop triggered by uncertainty or low confidence in perception may draw attention to conditions that would otherwise remain unnoticed, such as unstable environments, insufficient sensor coverage, or unanticipated edge cases in system design. Here, the anomaly functions as a signal that prompts inspection, discussion, or preventive action. Its value lies in making latent vulnerabilities visible, rather than in accurately identifying a specific problem.

A second form relates to learning and system improvement over time. Repeated anomalies—such as systematic misclassifications or recurring navigation errors—can expose structural limitations in models, training data, or environmental assumptions. In agricultural settings (FarmGO), such patterns reveal blind spots in detection logic; in simulation or maritime contexts, unexpected system behaviour can expose mismatches between modelled and actual dynamics.

Another form involves redundancy and distributed verification, particularly evident in safety-critical environments such as maritime navigation. Here, anomalies arise not only as system failures but as discrepancies between independent sources of information—such as instruments, charts, and visual observation. Rather than being treated as problems to eliminate, these discrepancies are operationally productive because they trigger cross-checking and reinforce situational awareness. In this configuration, deviation is not a breakdown of reliability but a mechanism for producing it through redundancy.

Taken together, these different forms of robot-perceived anomalies show that deviation within system logic is not uniformly problematic or beneficial. Its practical consequences depend on how it interacts with human interpretation, organisational routines, and environmental complexity. Anomalies therefore cannot be understood solely as technical irregularities, but must be seen as events that reorganise work, attention, and responsibility in socio-technical systems.

Human-perceived anomalies

In contrast to system-centred interpretations, human-perceived anomalies emerge from situated interpretation, embodied perception, and practical expectations about how work should unfold in concrete settings. Here, anomalies are not defined by statistical deviation or formal system thresholds, but by disruptions in meaningful activity, coordination, and shared understanding. What counts as an anomaly is therefore not a property of the system itself, but of how events are experienced, interpreted, and acted upon within specific social and material contexts.

Negative anomalies

A central category of human-perceived anomalies consists of situations where the system fails to register or respond adequately to conditions that humans immediately perceive as problematic or dangerous. The broken glass episode in the vignette is a paradigmatic example: for the robot, the situation does not necessarily constitute a detectable deviation, yet for staff it is an immediate safety-critical interruption requiring rapid intervention. These are blind anomalies from the system's perspective but highly salient from the human one. In such cases, the anomaly is not only recognised, but acted upon through urgent repair of safety and coordination.

When these situations occur, the consequences are typically negative in practical terms. Workflows are interrupted, attention is diverted from planned tasks, and human actors must compensate for the system's lack of situational awareness. Similarly, cases of mislocalisation or semantically incoherent robot behaviour—such as a system persistently interacting with a glass wall as if it were a door—produce breakdowns in intelligibility and coordination. Even when the system continues to operate according to its internal logic, the result is a disruption of shared situational meaning, which forces humans into corrective and supervisory roles.

A closely related form of negative anomaly concerns embodied safety risks and spatial misalignment. In environments such as restaurants, factories, or care facilities, humans rely on immediate perception of material conditions to ensure safety and continuity. When robotic systems fail to recognise hazards, obstacles, or relevant changes in the environment, the anomaly is experienced not as an abstract error but as a concrete and embodied risk. These events are particularly significant because they directly challenge the assumption of safe co-presence between humans and machines, requiring immediate intervention to restore stability.

Positive anomalies

At the same time, human-perceived anomalies are not exclusively negative. A significant part of the empirical material shows how deviations become productive through everyday improvisation, adaptation, and repair. In these cases, anomalies are not simply disruptions to be eliminated, but events that generate new forms of practice and understanding.

A first and highly prevalent form concerns everyday improvisation and repair work. Across settings, humans respond to unexpected system behaviour through local and often informal interventions: restarting systems, manually correcting errors, guiding robots, adjusting workflows, or bypassing malfunctioning components. Although these actions are rarely formalised or anticipated in system design, they are essential for maintaining operational continuity. At MedCare, a transport lane sometimes fails to start after trolleys have been placed. The worker solves it by gently kicking the area near the sensor: he does not know why it works, but "if it works, we do it." Such actions sit at the boundary between disturbance and repair: they may register as an external physical interference from the system point of view, but function as a stabilising technique in practice. (Report III, p.51). This

is where “positive anomalies” become visible as labour: the deviation is not an exception to be eliminated, but part of the informal maintenance regime that keeps the system usable. Over time, such interventions may become routinised, meaning that what initially appears as a deviation gradually stabilises into an accepted part of normal practice. In this sense, anomaly management becomes an embedded dimension of everyday work rather than an exceptional response to failure.

A second form involves tacit knowledge and embodied expertise. Human responses to anomalies often draw on forms of knowledge that are not fully codified or transferable into formal system specifications. Farmers interpreting subtle environmental cues, maritime operators relying on embodied navigation sense, or workers recognising early signs of machine instability all illustrate how situated expertise becomes visible precisely in moments of deviation.

Positive anomalies therefore expose the limits of formalisation: they reveal that not all relevant knowledge can be translated into data, rules, or models, and that effective system operation depends on forms of understanding that remain distributed, embodied, and context-dependent.

A third form concerns off-label system use and functional expansion. In many cases, users discover ways of working with robotic systems that were not anticipated in design specifications. Service robots may be repurposed for guidance or logistics support, industrial systems may be used in modified workflows, and agricultural or logistical technologies may be integrated into practices that differ significantly from their intended use. These adaptations can improve efficiency, reduce friction, or enable new forms of coordination. Importantly, they are often not recognised as “correct” uses from a design perspective, even though they may be highly effective in practice.

Taken together, these positive anomalies demonstrate that deviation is not only a source of disruption but also a mechanism through which systems are stabilised, adapted, and extended in real-world use. They show that human engagement with robotic systems is not limited to correction or mitigation, but also includes creative reconfiguration and continuous refinement of how systems are made to work in practice.

This relational understanding of valence and anomaly also provides a necessary bridge to the question of self-adaptivity in robotic systems. If anomalies are not fixed signals with stable meaning, but context-dependent deviations whose significance is produced in situated practice, then the premise of adaptation becomes more complex. Self-adaptive systems rely on the assumption that anomalies can be reliably detected, classified, and translated into meaningful system change. However, as the empirical material shows, what counts as an anomaly, and what it means in practice, is not self-evident. The next sections is therefore a discussion of what this understanding of anomalies means for accountable self-adapting robots.

Discussion of Anomalies in Self-Adaptive Robotic Systems

The distinction between negative and positive anomalies becomes particularly important when considered in relation to self-adaptivity. If autonomous systems are expected to adapt based on detected deviations, then anomalies are not peripheral disturbances but core mechanisms through which adaptation is initiated and sustained. However, the empirical material shows that this assumption is far from straightforward in practice.

A first issue concerns what counts as an anomaly in the first place. Across the cases, many relevant disruptions are never registered by the system at all (blind anomalies), while others are produced too readily through false positives. This creates a structural imbalance in the basis for adaptation: systems can only adapt to what they are able to detect, yet detection itself is unevenly aligned with what matters in practice. An example of this is the misreading of the environment in the HUMANA case in Report II, where a robot mistakes a glass wall for a closed door – which it has already passed. On a tour, HUMANA repeated: "Please open the door so I can go through," even though it was positioned in front of a glass wall rather than the open door it had just passed. Jonas eventually pressed the panic button. He explained that it was not unusual for the robot to misread points in the floor or ceiling. From the robot's point of view, the "anomaly" may represent a missing input or an assumed obstruction; from the human point of view, the anomaly is seen as a mis-localisation and a meaningless interaction with the environment (Report II, p.18). In engineering terms, this resembles "anomalies restricted to state": something matters in the world but is not observable to the agent through its sensory inputs. (Moustakidis, Passalis & Tefas, 2024, p.10).

A second issue concerns interpretive authority over anomalies. In practice, anomalies are frequently first identified, interpreted, and resolved by human actors rather than by the system itself. Human judgement determines whether a deviation is meaningful, how urgent it is, and what response is appropriate. This means that adaptation is rarely fully self-contained. Instead, it is distributed across human–machine interaction, where human intervention often stabilises or redirects system behaviour before any formal adaptive mechanism is activated. Self-adaptivity therefore operates as a hybrid process rather than an autonomous computational loop.

A third issue concerns the status of anomalies as learning signals. While self-adaptive systems are designed to treat anomalies as inputs for adjustment, the empirical cases show that many of the most consequential adjustments emerge outside formal learning pipelines. Everyday interventions—such as resets, manual corrections, bypassing failures, or informal workarounds—often stabilise systems effectively, yet remain invisible to adaptive algorithms. This produces a gap between operational adaptation (what humans do in practice to keep systems functioning) and computational adaptation (what systems formally register and learn from).

A fourth issue concerns responsibility under adaptive conditions. When systems change behaviour in response to detected deviations, responsibility becomes increasingly distributed and difficult to locate. It becomes unclear whether outcomes should be attributed to system design, real-time detection, human intervention, or the evolving interaction between them. Self-adaptivity does not remove responsibility; instead, it redistributes and sometimes obscures it across the socio-technical system.

Overall, these issues suggest that self-adaptivity should not be understood as a purely technical property of systems. It is an emergent and negotiated process shaped by the interaction between system detection, human interpretation, and organisational response. Across the material, the relationship between robot-perceived and human-perceived anomalies, as well as their negative and positive valence, further reinforces this point. The same event may be invisible in one register and critical in another, or simultaneously disruptive in one dimension and productive in another. A false positive may interrupt work and increase labour, while a missed hazard may expose a critical gap in system perception. Likewise, a system stop may be experienced as failure in the moment but later become a resource for learning and redesign.

This shows that valence is not fixed at the level of detection. Instead, it is produced through the interaction between registers and contexts of use. What appears as noise in one system may become meaningful signal in another, and what appears as failure may, over time, be reinterpreted as a resource. Understanding anomalies in this way also reframes their role in autonomy. Rather than representing exceptions to be eliminated, they might become constitutive of how autonomy is practically sustained.

Conclusion on Anomalies

Across the empirical cases, anomalies are not exceptional breakdowns or rare edge cases, but constitutive elements of HRI in practice. They appear continuously in the everyday operation of systems across restaurants, farms, industrial environments, and maritime settings, shaping how work is organised, how decisions are made, and how systems are kept functioning under real-world conditions.

When analysed through valence, anomalies reveal a fundamental dual dynamic. On the one hand, they disrupt workflows, expose the fragility of computational models, and generate uncertainty, additional labour, and moments of breakdown. These negative manifestations are often what make system limitations visible in the first place, forcing immediate intervention and revealing mismatches between system expectations and environmental complexity. On the other hand, the same kinds of deviations frequently become the basis for repair, learning, and adaptation. They trigger human improvisation, generate local solutions, and provide the empirical material through which systems are gradually adjusted, refined, or stabilised.

Our empirical research shows that anomalies are relational phenomena whose meaning and consequence emerge only through situated interpretation and practical engagement. What counts as an anomaly, and whether it is experienced as problematic or productive, depends on how it is perceived within a given context, how it is acted upon by human and system actors, and how it is absorbed into ongoing organisational routines. The same event may simultaneously interrupt one process while enabling another, or be treated as a failure in one moment and as a resource in the next.

This matters directly for RoboSAPIENS' ambition to use anomalies as triggers for adaptation (Birch et al 2026): if only anomalies "set the technology in motion," then it becomes a core problem when the system's anomaly signals are either (a) missing where humans see critical anomalies, or (b) overly frequent and costly due to false positives. (Report IV, pp.4–5; Report II, p.8).

Importantly, this has direct implications for the design and understanding of autonomous and adaptive robotics. If anomalies are to function as triggers for adaptation, then the entire chain from detection to interpretation to response becomes a central design concern. Taking the wider view then adaptation does not emerge from anomaly detection alone, but from the broader understanding that determines how and why deviations are made actionable. In this way anomalies structure the conditions under which systems remain operational, intelligible, and governable. In practice anomalies are absorbed, corrected, ignored, reinterpreted, or transformed into new routines, often in ways that are only partially visible to formal system design.

Sim2Real

"The map is supposed to function as a 1:1 simulation of the warehouse, and the AGVs navigate solely based on their position on the map. This can lead to problems if the AGVs think they are in one place but are actually somewhere else. For example, this can happen if the floor is too slippery or the wheels are too worn, so that their braking distance becomes longer than normal. Over time, this can cause them to drift out of sync with the map and eventually cause them to stop driving at all. Therefore, the maps are also synchronised and updated continuously to avoid such errors." (Report III, pp. 36-37)

[Introductory vignette drawing on Report III illustrating several aspects of Sim2Real dealt with below.]

Introduction

The previous chapter showed that anomalies are not rare interruptions at the edge of otherwise smooth robotic operation. Rather, they are relational events in which the practical limits of autonomy become visible, because "normal" is negotiated across models, environments, and human responsibilities. In this sense, anomalies reveal not only that systems fail to register, classify, or respond as expected, but also that the worlds robots operate in exceed what their internal models can fully anticipate. The question that follows, then, is why these misalignments persist so systematically across settings. What is it about robotic systems that makes anomalies such a recurring condition of real-world use rather than an occasional exception?

Sim2Real takes the next analytical step by addressing precisely that question. If the anomalies chapter examined moments where models, environments, and practical expectations fall out of alignment, Sim2Real examines the conditions that contribute in producing such misalignment. Robotic systems do not encounter the world directly. They operate through simulations, maps, routes, internal models, and other formalised representations of the environments in which they are expected to act. These representations are indispensable, but they are always partial. They make robotic action possible, yet they can never fully eliminate the gap between model and world. Sim2Real therefore helps explain why safety must remain revisable, why trust in data and models remains fragile, and why anomalies keep recurring in practice.

Modelling and simulation is an established phase in the engineering of many robotic systems. Broadly speaking simulation functioning here as a virtual model in which an environment is computationally constructed and maintained. It is typically employed to assist, on the one side with the design and testing of robotic systems, and on the other as an integral part of the run-time logic used to assist them during real-world operation. In other words, it underpins automation by enabling developers to model physical processes, control strategies, and system interactions in a controlled, repeatable, and cost-effective manner (Barricelli & Fogli, 2024). Simulation operates as an intermediary layer between programmed control logic and real-world robotic performance, aligning model parameters with physical phenomena in order to support predictable and transferable behaviour. It is this alignment between the simulated and the real-world settings that we have referred to during our empirical work as Sim2Real. While engineering perspectives

often frame simulation as a means of achieving accuracy, prediction, and control, the practical consequence as illustrated by the above extract reveals a more complex picture in which a Sim2Real gap between simulation and the actual environment becomes evident.

Specifically the introductory vignette highlights the dynamic and ever-changing nature of the real world, emphasizing that simulations are not static technologies capable of fully capturing the many complexities of the warehouse setting. However, the extract also points out that unpredictability does not only stem from the external environment, but even from the robot itself. Over time, robots can deteriorate, and this gradual wear can transform them into liabilities rather than assets. In the case of the vignette, it is the robot's wheels that become worn. The wheels became something the employees closely monitored and by drawing on experience, they tried to predict wear before it caused problems, allowing them to intervene early and reduce the consequences. These employees had developed an awareness of the robots' physical limitations, such as recognizing when the wheels were becoming worn down. They knew what signs to look for and could anticipate when these components would begin to compromise performance. This practical, experience-based knowledge allowed them to intervene before failures occurred, something that the simulation alone could not fully account for.

Seen from a technical perspective, the Sim2Real gap can be attributed to several interrelated factors. First, in models there is the fundamental difficulty of achieving precise alignment between simulated and real-world conditions, given the need for abstraction and simplification in any computational representation. Second, achieving real-time data exchange between physical systems and their simulated counterparts may be difficult to fully realized in practice due to issues of real-time data exchange without partial, delayed, or selectively interpreted data transfer. Third, robotic systems as seen here to operate in environments that exceed what can be exhaustively modelled, because they are dynamic, context-dependent, and socially structured. With a focus on the human interaction with robotic systems, it is this latter limitation, the difficulty of modelling a situated social world, that becomes especially significant to this section of the report. Accordingly, and in line with the broader themes of this report, the Sim2Real gap is not treated here solely as a technical problem of representational accuracy. Rather the simulation is seen as part of a socio-technical arrangement in which alignment between simulation and world is in part actively produced by human actors.

Exploring Sim2Real through the empirical material directs attention to the intersection between the technical capacities of simulation such as representation, prediction, and coordination, and the forms of human behaviour and response that emerge in real-world settings. This analysis shows that the effectiveness of constructing and maintaining virtual representations of the world does not depend solely on technical fidelity, or on how accurately particular aspects of reality are modelled. Rather, it also depends on the ongoing engagement of human actors who interpret outputs, adjust parameters, and actively sustain the relationship between simulated and real environments in practice.

The Sim2Real Gap

The existence of a gap between simulation and the reality it seeks to replicate is not a new insight; it is widely recognized among engineers and programmers. However, what this gap means in practice, and particularly for the humans who work with robots developed through simulation, is what our fieldwork set out to explore.

A central limitation of simulation is its tendency to simplify the world. While it can capture many aspects of an environment, reality remains far more complex and nuanced than any model can fully represent. Across our cases, a recurring pattern emerges: simulations account for a large portion of the environment "well enough," but the remaining share (often described as the final 20%) is where accidents, surprises, and innovation occur.

At FarmGO (Report IV, p. 12), this limitation is explicitly acknowledged. Their simulations are not fully realistic: vine heights vary, soil conditions shift, and weather and visibility constantly change. Simulations are therefore used to establish sensor placement and basic functionality, making systems "good enough" before deployment. The most significant learning happens in the field, where robots encounter anomalies such as incorrect tags, unexpected obstacles, or wildlife.

A similar pattern appears in service robotics (Report II, p. 11) where challenges arise precisely in situations that are difficult to model: chairs being moved, children running unpredictably, customers stopping suddenly, or changes in lighting and floor conditions. While these situations are entirely ordinary from an ethnographic perspective, they are often treated as edge cases in design.

In cobot and AGV deployments (Report III), even small deviations, such as a misplaced pallet, a trolley parked incorrectly, or a half-open door, can disrupt carefully planned routes and safety systems. What emerges from the ethnography is that these "last 20%" are not rare exceptions but rather the fabric of everyday life. Simulations, however, tend to formalize environments as though they more closely resemble the stable 80% than they actually do.

The key insight across cases is not that simulations are ineffective. Rather, they are highly effective at preparing systems and people for the expected, general, and average conditions of operation. It is in the remaining, messy, and unpredictable aspects of reality that their limitations become visible, and where human adaptability becomes essential.

Technologies used in automation

Throughout this report, many different technologies are mentioned in relation to simulation. To help the reader, we provide a brief overview of these tools. It should be noted that, as anthropologists, we are not experts in these technologies; this overview is based on our empirical observations, interviews, and supplementary internet research. The technologies are not simulation technologies, but can help robots navigate in spaces. While there are certainly many other technologies available, these are the ones most relevant to the cases we studied.

LiDAR: A sensor system that uses laser light to measure distances and create detailed 3D maps of environments.

RFID: Radio-frequency identification; tags and readers track objects or components for navigation, inventory, or workflow simulations.

CAE: Computer-Aided Engineering; software used to simulate and analyze physical systems.

Visual SLAM: Simultaneous Localization and Mapping using cameras; enables robots to map environments and track their position in real time.

CAD: Computer-Aided Design; software for creating precise 2D or 3D models of objects, machines, or environments for simulation purposes.

ECDIS: Electronic Chart Display and Information System; digital navigation system for ships that can be simulated to train crews and plan routes.

SPOS: Ship Performance Optimization System; software that simulates vessel performance and routing to improve efficiency and safety.

We were particularly impressed by the accuracy and sophistication of the simulation tools. As Nick at ElectroFlight explained, the software is "basically like Excel or Word" but its effectiveness depends entirely on the expertise of the user and the quality of the input data (Report III, p. 14). In other words, even the most advanced programs can only be as relevant and accurate as the people operating them. This underscores an important point that runs throughout our cases: simulation based on 'input' technology is not a fully autonomous solution, but a tool whose value emerges through human interpretation, input, and oversight.

At MakeBot, employees described how their robots use sensors placed around the body to navigate the space with remarkable precision, almost like dancing through the environment (Report III, p. 26). This example illustrates a broader pattern: across all our cases, robots rely heavily on underlying technologies - sensors, cameras, LiDAR, visual SLAM, and other navigation tools - to translate robot movements into effective real-world operation. Crucially, the effectiveness of these systems is contingent on the integrity of the data and the environment: most accidents or anomalies occur when a sensor is blocked, damaged, or misaligned.

Taken together, these observations highlight a key theme of this report: while simulations can model environments, tasks, and machine behaviors with impressive accuracy based on other technologies such as LIDARs, their usefulness is inseparable from human knowledge, interpretation, and intervention. Simulations provide a framework or reference, but it is the combination of technology and human judgment that allows robots to function reliably in the complex, dynamic realities of workplaces, farms, ships, and warehouses.

Themes in the empirical data

Across our material, Sim2Real is never just a technical issue. It is about the gap between lived, bodily experienced environments and the cleaned-up, parameterised versions of those environments that live in code, maps, and simulations. Looking across the service robots of Report II, the cobots and industrial settings of Report III, and the dynamic environments of fields and seas in Report IV, a consistent pattern emerges. Simulations are powerful and often highly useful tools, but they are always partial. They may capture much of an environment well enough to support design, training, or deployment, yet the parts left out are often precisely where the most important frictions, risks, and adjustments appear in practice.

Across cases, four broader findings stand out. First, simulations reduce complexity, but they do so by stripping away many of the variable and situated conditions that matter in lived environments. Second, simulations are useful not only because they model expected conditions, but also because they can reveal errors, test assumptions, and support

diagnosis. Third, simulations remain workable only because human actors continuously align model and world in practice through observation, intervention, and interpretation. Fourth, simulations are never neutral descriptions of reality. They always reflect selective perspectives on what matters, what counts as risk, and whose knowledge is treated as relevant.

1. Simulation reduces complexity but cannot keep pace with lived environments

A central cross-case finding is that simulation works by reducing the variation of real environments into forms that can be represented, computed, and acted upon. In almost every case, simulation was used to tame an environment by stripping away complexity. Service robots were developed and tested against tidy floor plans with straight aisles, stable table layouts, fixed charging spots, and relatively predictable human movement (Report II). Cobots and AGVs were designed around stable work cells, known pallet sizes, repeatable sequences, and safety zones (Report III). Even where no explicit simulation software was foregrounded, there was an implicit model of an ordered environment in which deviations would be relatively rare. In this sense, simulation makes robotic action possible by creating a manageable world. But it also does so by leaving out many of the features that characterise ordinary life as it is actually lived.

This limitation became especially visible in cases where robots depended on relatively fixed maps and routes. At Restaurant 1, staff explained that tables could not be freely rearranged without disrupting the robot's navigation. At Restaurant 2, multi-robot coordination worked in principle, but became fragile once it had to unfold in a shared space with humans moving unpredictably (Report II, p. 11). At MEASSURE and HUMANA, preprogrammed routes made systems vulnerable to everyday changes and small disruptions (Report II, pp. 21, 25). At HandsON, AGVs navigated according to a near 1:1 digital map of the warehouse, yet changes such as slippery floors or worn wheels could gradually throw the vehicles out of sync with that representation (Report III, p. 40). Across these cases, the issue was not that the systems failed to do what they were designed to do. It was that the real environments in which they operated remained more fluid, shifting, and contingent than the simulation could continuously keep pace with.

The same point appears in agriculture and maritime contexts, though in even sharper form. At FarmGO, simulations were explicitly described as useful but necessarily incomplete, since uncertainties such as vine height, soil conditions, weather, visibility, wildlife, and field variability could never be exhaustively captured in advance (Report IV, p. 12). In maritime contexts, the same applies to sea state, ice conditions, changing traffic, and uncertain data (Report IV, pp. 23-24). Across the cases, the practical lesson is not that simulations are ineffective. Rather, they are highly effective for the expected, average, and general aspects of operation. Their limits become visible in the remaining, messy, and context-dependent aspects of reality — not because these are rare, but because they are the fabric of everyday life.

2. Simulation is useful because it prepares, reveals, and diagnoses

If the first finding is that simulations simplify reality, the second is that they remain indispensable precisely because they make certain forms of preparation, testing, and discovery possible. Across our cases, practitioners placed considerable value on simulation. FarmGO treated simulation pragmatically as better than having no model at all (Report IV, p. 12). ShipGO instructors showed genuine enthusiasm for simulators as pedagogical tools (Report IV, p. 16). ElectroFlight used simulation to begin development before hardware was available, thereby reducing time and cost (Report III, pp. 13-14).

In this sense, simulation is not valuable because it reproduces reality perfectly, but because it allows engineers, operators, and trainees to rehearse scenarios, test configurations, and identify likely problems in advance.

The empirical material also shows that simulations can reveal hidden features of the real world rather than simply approximate it. At ShipGO, a discrepancy in the simulator initially appeared to be a modelling flaw but turned out to originate in the actual design drawings of the vessel (Report IV, p. 23). At ElectroFlight, simulations supported the modelling of customer tools and tasks before installation, reducing constant physical adjustment (Report III, p. 13-14). At FarmGO, simulation helped determine the placement of cameras and sensors before deployment, even though the final 20 percent of field conditions still had to be learned in practice (Report IV, p. 12). In these cases, simulations do not eliminate the gap between model and reality. They make that gap visible in productive ways, allowing practitioners to identify assumptions, inconsistencies, and practical constraints before full deployment.

This is also the place where a much shorter version of the current "Imagined Future" section can be retained. Simulations do not only model what is currently possible; they also help shape expectations of what might become possible. But in the empirical material, the strongest point is not only the long cultural history of automata. It is the present-day use of simulation as a tool for managing expectations in concrete development work. As Nick at ElectroFlight described, simulation can be used not only to imagine desired functionalities, but also to demonstrate limitations and show why certain tasks are not feasible (Report III, pp. 13-14). In that sense, simulation works in two directions: it can inspire future possibilities, but it can also ground expectations in current technical realities. That is the part of the theme that belongs here.

3. Humans continuously align model and world

The third major finding is that the usefulness of simulation depends on continuous human effort. Across the cases, humans do not merely use simulation-based systems after design is complete. They actively sustain the relation between model and world through observation, adjustment, repair, and interpretation. At the restaurants, staff had to reposition robots so they could see white ceiling dots again after losing orientation (Report II, p. 15). In service settings more broadly, staff learned through experience which layouts, corners, and rhythms of movement created trouble, and they guided robots with gestures, furniture arrangements, and social coordination. At HandsON, workers noticed worn wheels before the AGVs stopped functioning, and they learned that high shelves could confuse robot perception (Report III, pp. 40, 37). In these cases, simulation remains workable only because humans compensate for its limits in practice.

This human work is not limited to minor fixes. It also includes forms of tacit and embodied expertise that cannot easily be formalised. In maritime contexts, instructors deliberately inserted ECDIS failures into simulations so students could learn to revert to traditional navigation and look outside rather than over-rely on digital representations. Experienced captains develop embodied familiarity with ice conditions, sound, movement, and environmental cues that are extremely difficult to encode for robots. Similarly, in agriculture, concern for fawns in the field reflects forms of practical and ethical knowledge that exceed what hazard zones or sensors can neatly represent. These examples of what is commonly called "fault injection", show that Sim2Real is not simply bridged by more data. It is navigated through forms of judgment, experience, and bodily learning that remain partly outside the simulation itself.

A particularly important point is that humans do not only compensate for simulation gaps after the fact; they also probe them and feed back into design. At ElectroFlight, observation of workers placing thin metal plates revealed a subtle "dipping" motion that would have been difficult to infer from task descriptions alone (Report III, p. 55). Once understood, this embodied technique was incorporated into the robot's programming. From a Sim2Real perspective, this is one of the clearest examples of how human practice becomes a diagnostic resource. Workers do not merely reveal where the model fails; they also reveal which features of real practice should have been in the model in the first place. Across the cases, Sim2Real works best when anomalies, workarounds, and local experience are not dismissed as noise, but treated as inputs for ongoing refinement.

4. Simulations embed selective worlds

A fourth cross-case finding is that simulations are never neutral. They do not simply copy reality; they formalise a selective version of it. Every simulation reflects choices about what matters, what counts as a risk, what is treated as relevant signal, and what is left out as noise. This means that Sim2Real is not only about whether the model is accurate enough, but also about whose world the model represents. That is why the problem cannot be reduced to a narrow issue of technical fidelity. It is also a question of perspective and priority.

The cases make this visible in different ways. In FarmGO, simulations initially reflected standards, machine-builder priorities, and formal risk categories, while farmers' more situated concerns — such as fawns in the field or local knowledge of terrain — entered later and often only indirectly (Report IV, p. 12). In maritime contexts, systems such as ECDIS and SPOS formalised hydrographic data, commercial priorities, and regulatory assumptions, while lived seamanship entered more as a corrective than as a design premise (Report IV, pp. 17, 22). In service and cobot settings, vendor simulations privileged smooth workflows, stable layouts, and compliant users, leaving the negotiated realities of public and workplace life underrepresented. At MedCare, MES and robot logic formalised a particular view of what counted as a normal process, while staff developed workarounds and parallel routines to bridge the gap between that formal model and the demands of practice (Report III, pp. 44-45). Across these examples, simulations appear not as transparent mirrors of the world, but as selective socio-technical constructions.

This is also why local and embodied knowledge matters so much. When simulation becomes the reference world for design and planning, low-status, practical, or tacit knowledge can easily be treated as secondary. Yet the fieldwork shows that systems function best when such knowledge has a route to push back into formal representation: when farmers' concerns shape sensor placement, when captains influence scenarios, when staff experiences reshape workflows, and when workers' tinkering informs later adjustments. Sim2Real therefore concerns not only alignment between model and world, but the social conditions under which certain parts of the world become modelled at all.

Discussion of Sim2Real in self-adaptive robotic systems

Simulation and modelling are central to self-adaptive robotic systems such as those envisioned in RoboSAPIENS, because they provide the system with a way of reasoning about action before acting in the world. In broad terms, data gathered

through monitoring can update internal models of the robot and its environment; these models can then be used in analysis to detect deviations or anomalies, and in planning to explore possible courses of action through "what-if" scenarios. In this sense, simulation supports self-adaptation by allowing the system to anticipate outcomes rather than merely react after the fact. Within MAPLE-K-like architectures, simulation may therefore function both as part of the system's knowledge base and as a means of ongoing adaptation. Yet the argument developed in this chapter suggests that such simulation cannot be treated as a self-sufficient solution. Its effectiveness depends fundamentally on the degree to which the simulated world remains aligned with the realities in which the system must actually operate.

At the core of this challenge is the fact that robotic systems never encounter the world directly. They operate through models, maps, and representations that are necessarily partial, selective, and shaped by prior assumptions. While self-adaptive systems promise to reduce this gap through continuous updating, the cases show that real-world environments are not simply dynamic in a technical sense. They are also context-dependent, socially organised, and interpreted through human practice. Conditions such as worn components, shifting layouts, ambiguous objects, environmental variability, or changing workflows are not just missing variables waiting to be incorporated into more accurate models. They are expressions of environments whose relevance is continuously negotiated by human actors.

This could have important implications for self-adaptivity, since rather than operating as closed loops in which systems detect, analyse, and correct deviations autonomously, adaptive processes are in practice distributed across human–machine interaction. Many of the adjustments that keep systems aligned with their environment emerge through human interpretation, intervention, and ongoing calibration. People reposition robots, compensate for system drift, reinterpret outputs, and develop local strategies that stabilise operation in ways that are not fully captured by formal models. In this sense, alignment between simulation and reality is not achieved once and for all through design or runtime updating but is continuously sustained through a combination of technical representation and situated human practice.

This also challenges the assumption that increasing model fidelity or reducing human involvement necessarily leads to more robust autonomy. Systems that are highly detailed in their representation of certain variables may still fail to capture the social and practical conditions in which they operate. Conversely, systems with limited representational fidelity may remain workable because human actors actively compensate for their shortcomings. The empirical material suggests that the most robust forms of alignment are not those that eliminate human involvement, but those that allow human knowledge, experience, and judgement to feed back into modelling, interpretation, and adjustment over time.

For self-adaptive robotics, this raises a new question: not only how systems update their models, but how these updates remain meaningful and interpretable within the contexts in which they are used. If adaptive systems continuously modify routes, thresholds, classifications, or behaviours, users must still be able to understand when and why these changes occur, and how to respond to them. Without such interpretability, adaptation might risk introducing new forms of opacity, where the system's representation of the world evolves in ways that are no longer intelligible to those who rely on it. In such cases, adaptation may intensify rather than resolve the gap between model and world.

Seen in this way, Sim2Real is not a technical problem to be solved once sufficient data or computational power is available. It is an ongoing condition of operation in which models remain necessarily incomplete and must be continuously aligned with lived reality. Self-adaptivity does not remove this condition but makes it more dynamic, as both the system and its representations change over time. The questions might therefore be how to improve models, but how to organise adaptive systems in ways that remain responsive to situated knowledge, and capable of supporting meaningful human engagement.

At the same time, this chapter also points directly forward to Sabotage/Tinkering. If Sim2Real names the persistent gap between formal representation and lived reality, then the next chapter shows what people actually do when that gap becomes consequential in practice. They probe, adjust, workaround, bypass, kick, re-route, test, and sometimes resist the system outright. As the introduction to Sabotage/Tinkering already makes clear, these practices are not marginal to robotics, but central modes of engagement through which systems are integrated, negotiated, and made useful in everyday settings. The same examples that appear in the Sim2Real chapter as human compensation and alignment work reappear there as tinkering, and sometimes as sabotage, depending on intention, context, and consequence. Sim2Real therefore prepares the ground for that chapter by showing why such interventions arise in the first place: not simply because humans interfere with otherwise complete systems, but because models remain partial and must be supplemented, corrected, or contested in practice.

Conclusion on Sim2Real

Taken together, this chapter has shown that Sim2Real is not simply a technical problem of transfer from simulation to deployment. It is a structural condition of robotics in the wild. Robotic systems do not act on the world directly, but through maps, routes, internal models, and simulations that make action possible by reducing complexity into calculable form. These representations are indispensable, yet they are always partial. They can support design, training, coordination, and runtime reasoning, but they cannot fully capture the shifting, messy, and socially organised environments in which robots must operate. The question, then, is whether Sim2Real can meaningfully be treated as a problem to be solved at all, or whether it is better understood as a persistent condition of operation. From this perspective, the gap between simulation and reality may not be an occasional imperfection that improved modelling will eventually eliminate, but an inherent feature of systems that rely on representation. If so, autonomy would not converge toward full independence from this gap, but remain distributed, negotiated, and continuously supported in practice

The chapter has also shown that the Sim2Real gap does not simply produce failure. It also produces human work. Across the cases, operators, workers, instructors, farmers, and technicians interpret outputs, notice drift, adjust parameters, reposition robots, update maps, and feed situated experience back into formal models. Human engagement therefore appears not as an accidental supplement to simulation-based systems, but as one of the ways those systems remain operational at all. The empirical data in this chapter argues this: the strongest form of Sim2Real is not detached abstraction, nor purely data-rich automation, but a socio-technical arrangement in which technical fidelity and human

engagement support one another. Alignment between model and world is not pregiven in design; it is continuously produced through interpretation, intervention, and practical learning. Simulation, in this sense, redistributes rather than removes human judgment.

If self-adaptive robotics is to remain safe and trustworthy under unknown change, then the question is not only how internal models can be updated, but how those updates remain meaningful in environments that are socially and materially more complex than any model can fully anticipate. Self-adaptation may reduce some forms of Sim2Real mismatch, but it cannot abolish the need for interpretation, oversight, and practical correction. A more realistic ambition is therefore not to eliminate the gap between model and world, but to design systems that can navigate it more reflexively: systems in which anomalies become visible, human knowledge can feed back into modelling, and adaptation remains legible enough to support reliance rather than undermine it.

Sabotage/Tinkering

"They also explore their own abilities by thinking beyond the existing technologies and in a way innovating on them. "Tinkering" thus helps employees expand their knowledge about the robot, the technologies and their own abilities. However, this is seen as sabotage in relation to the robot and from developers like ElectroFlight and is therefore something that must be corrected. The question is, though, whether it is even possible to avoid this "tinkering" or whether humans will always find a way to experiment with the technology. We asked Nick about the motivations behind the sabotage examples and whether he thinks it is because employees do not trust the technology. He replies: "They probably can't fully grasp what it means. So, they think we can just… It usually just runs there, so we can just manage to walk over here." He thus has the perception that employees do not understand the seriousness of the robot technology and therefore carry out this sabotage/tinkering. "(Report III, p. 12)

[Introductory vignette drawing on Report III illustrating several aspects of sabotage and tinkering dealt with below.]

Introduction

Sabotage has become an increasingly recognised concept within robotics and human–robot interaction (HRI), but the vignette above already suggests why it cannot be understood simply as humans obstructing machines. What appears from an engineering perspective as misuse, interference, or even sabotage may in practice also involve experimentation, frustration, local problem-solving, or attempts to reclaim a measure of control over opaque systems. During our fieldwork, paying closer attention to such actions revealed something important: sabotage and tinkering are not merely side effects of robotic deployment. They are among the ways robotic systems are actually encountered, negotiated, resisted, and made usable in everyday life. Understanding them is therefore not only a matter of prevention. It is also a way of learning where systems remain rigid, fragile, mistrusted, or socially contested.

Across Reports II, III, and IV, sabotage and tinkering emerged not as isolated incidents but as recurring modes of engagement with robotic systems. Rather than treating them as wholly separate categories, this chapter understands

them as points along a spectrum. At one end lies tinkering: playful, curious, exploratory, or practically corrective interaction through which people test the system, learn its limits, and try to make it work in context. At the other end lies sabotage: actions carried out with the intention to obstruct, undermine, or discredit the robot's operation. Between these poles lie many ambiguous cases in which the same act, like stepping in front of a robot, overriding a stop, interfering with a route, or "helping" a system along, can be exploratory in one setting, corrective in another, and oppositional in a third. Intention matters, but it is not always directly visible. For that reason, sabotage and tinkering are best approached not as fixed behavioural types, but as relational positions within a broader field of engagement between humans and robots.

Although these behaviors can be placed along a spectrum, we argue that they should not be understood as fixed or stable states. Rather, they are fluid and can shift within the same interaction. A person may begin by curiously tinkering with a robot and, moments later, perhaps due to frustration or uncertainty, move toward obstructing or even damaging it. In this sense, the temporal dimension of tinkering and sabotage is crucial, as such transitions can occur rapidly and without clear boundaries.

Importantly, this chapter does not treat sabotage as a predefined analytical category. Instead, both "sabotage" and "tinkering" emerged inductively from the fieldwork, when informants explicitly described certain actions that way. Neither concept has been prominent within RoboSAPIENS research to date (in contrast to the other main categories such as Safety, Trust, Anomalies and Sim2Real). The concept of tinkering made it possible to revisit some of the cases explored empirically and ask whether they might instead be understood as exploratory or compensatory engagements. This shift is important because it moves the analysis away from a simple opposition between rule-following and rule-breaking. What matters more is why people engage robotic systems in unexpected ways, what such actions reveal about the system's fit with practice, and what they tell us about the practical limits of autonomy in the wild. In this sense, sabotage and tinkering are not isolated from the rest of the report. They are tightly bound up with the previous themes: they shape how safety is stretched and negotiated, how trust is built or weakened, how anomalies are handled, and how Sim2Real gaps are bridged, exposed, or contested in practice.

The distinction encountered through the safety course between "reasonably foreseeable misuse" and "unintended use" is useful here, but only up to a point. As Rosa explained, unintended use refers to forms of interaction that diverge from intended operation but do not necessarily place the system outside its safety envelope (Report III, p. 26). This category helps illuminate how industry frameworks already expect users to engage with robots in ways designers did not plan. Yet the empirical material also shows the limits of this language. A small kick that gets a stuck robot moving may count as unintended use, but it is analytically very different from hitting a robot out of frustration or obstructing it because it is experienced as a threat. Both may sit outside intended use, but they differ in motivation, meaning, and consequence. This is why the present chapter moves beyond the industry distinction. The key question is not only how users deviate from intended operation, but how and why they do so, whether to learn, to compensate, or to resist.

For that reason, the empirical section that follows is organised around three connected themes, each representing a different position on the spectrum of sabotage and tinkering. The first theme, learning the system, examines how users

test, probe, and experiment with robots in order to understand them. Here tinkering appears as part of adoption: people build familiarity, practical skill, and bodily trust through exploratory engagement. The second theme, making the system workable, turns to the many compensatory interventions through which people adjust, correct, and quietly repair robotic systems so they can function under real-world conditions. Here the key point is that users do not merely adopt systems as given; they often complete them through local workarounds and situated forms of design. The third theme, resistance, refusal, and legitimacy crises, examines what happens when interventions are no longer primarily about learning or helping the system along, but about contesting it. Here sabotage appears most clearly: not simply as destructive behaviour, but as a sign that the system is experienced as unfair, imposed, threatening, or misaligned with people's own sense of good work, responsibility, and autonomy. Together, these three themes allow the chapter to move from exploration, through compensation, to resistance, and thus to show more clearly how sabotage and tinkering are related without collapsing them into one another.

Seen in this way, sabotage and tinkering are diagnostic categories. Where they appear, something important is happening: a gap between promised and actual performance, between formal procedures and everyday work, or between technical control and workers' own sense of responsibility and identity. This is why they matter for the broader argument of Report V. If autonomy in the wild is not a fixed machine property but a distributed and negotiated accomplishment, then sabotage and tinkering show what happens when that accomplishment remains unstable. They reveal that autonomy is never only about what control has been moved into the robot. It is also about how much room humans have to test, reshape, and, when necessary, resist the systems that enter their workplaces, ships, farms, hospitals, and public spaces.

Themes in empirical data

As the above introduction clarified this chapter does not treat sabotage and tinkering simply as problems of unintended use. Instead, it asks what such actions reveal about robots in the wild. Across the cases, they reveal at least three recurring dynamics.

First, we explore how people are learning the system by testing, probing, and informal skill-building. Children step in front of restaurant robots to see whether they stop; visitors test how HUMANA reacts; workers experiment with buttons, interfaces, and routines in order to understand what a system does (Report II, Report III). In this sense, tinkering is not external to adoption, but part of how adoption happens. It is one of the ways people turn opaque technical systems into something they can orient themselves around bodily and practically.

Secondly, people are making the system workable through workaround, correction, and risky compensation. Sabotage and tinkering reveal the extent to which robotic systems often do not fully fit the realities into which they are introduced. Users do not only test robots out of curiosity; they also intervene because systems are too rigid, too slow, too opaque, or too poorly aligned with workflow. Workers reposition robots, bypass formal hierarchies, create small hacks, and develop local routines that make the technology workable in practice (Report III). In such cases, what may appear from a design perspective as deviation or interference can also be understood as practical correction. Tinkering,

here, becomes a form of situated completion of design: users do not simply adopt the system, but actively help make it function under real conditions.

Thirdly, the cases show that some interventions are not primarily exploratory or corrective, but tied to resistance, refusal, and legitimacy crises (Report II, Report III). Sabotage appears most clearly where robotic systems are experienced as threats to autonomy, competence, responsibility, or good work. Blocking sensors, placing objects in front of robots, ignoring or sidelining systems, or refusing their introduction altogether are not simply signs of ignorance or irrational dislike. They often point to deeper legitimacy problems: fears of replacement, dissatisfaction with imposed workflows, uncertainty about responsibility, or resistance to systems that challenge established professional judgment. In this sense, sabotage is analytically important not because it is always destructive, but because it shows where the social and organisational basis of robotic adoption begins to fracture.

1. Learning the system: testing, probing, and informal skill-building

A recurring pattern across our cases is that people do not simply encounter robots as finished and self-explanatory technologies. They learn them by testing them. Before robotic systems become routine tools, they are often met through bodily probing, playful experimentation, and informal attempts to understand what they can do, how they react, and where their limits lie. In this sense, tinkering is not external to adoption. It is one of the practical ways adoption happens. People do not only receive knowledge about robots through manuals, demonstrations, or formal training; they also acquire it by stepping in front of them, pressing buttons, trying things out, and observing how the system responds.

This pattern was especially visible in the service-robot cases. At the restaurant, one of the waiters explained that the robots were particularly entertaining for children, who often ended up “playing” with them by stepping in front of them to see whether they would stop, or by touching them as they passed (Report II, p. 17). Rather than interpreting such actions simply as obstruction, they can be understood as exploratory engagements through which users test the robot’s safety logic and begin to build a practical sense of how it works. The same was true in our own fieldwork experience. Even though we had been told, and had already observed, that the robot would stop when encountering obstacles, we still felt compelled to test this ourselves (Report II, p. 9). This bodily engagement mattered because it shaped our own sense of whether the robot was safe and how it behaved in practice. A similar dynamic appeared at HUMANA, where both we and other visitors stepped in front of the robot to see how it would react. As Jonas explained, this kind of behaviour was common among guests (Report II, p. 21). What is important in all these examples is that understanding does not arise from abstract information alone. It is produced through direct interaction, trial, and embodied experience.

These cases also show that such exploratory conduct is not limited to playful public interaction. In workplace settings, informal learning through tinkering can become a necessary part of developing competence around robotic systems. Martin’s account from MakeBot is particularly revealing. The first time he had to interact with the robot’s interface, he did not dare touch anything because he was afraid he might break something (Report III, p. 25). He had not received formal onboarding or training, but was simply told to go and “play” with the robot by pressing buttons. In principle, this

invitation positioned tinkering as a route to learning. In practice, however, Martin's lack of technical understanding made that invitation difficult to act on. He described the technology inside the robot as a "black box": something clearly active and consequential, but not yet intelligible to him (Report III, p. 25). The point here is important. Tinkering is not always easy or immediately empowering. It can also be hesitant and uneven, depending on how much prior knowledge a user has and how confident they feel in relation to the system. Yet even in Martin's case, the issue was not that learning through tinkering was irrelevant. It was that informal experimentation requires some minimum sense of safety and interpretability in order to become productive.

A similar insight emerged in the maritime material. At ShipGO sailors and captains explained that they often have to become acquainted with new technologies through direct use rather than through any standardised interface logic. As Curt put it, the first weeks on board a new ship are often spent "playing 'guess a new button'." (Report IV, p. 24). This formulation is striking because it makes explicit what is often implicit elsewhere in the material: learning a technological system frequently means experimenting with it in practice, not simply receiving a complete explanation in advance. Here too, tinkering functions as a way of turning unfamiliar technology into situated competence. Because there is no single standardised arrangement across ships, crews must learn each configuration through practical exploration on the job. This suggests that tinkering is not just something that happens when design fails. It is often built into the ordinary process through which complex technologies become workable for those who must use them.

Seen across these cases, informal skill-building through testing and probing is one of the clearest ways in which humans reclaim agency in relation to robotic systems. Robots often arrive with strong promises, technical authority, and partially opaque forms of operation. Tinkering allows users to bring such systems down to the level of practical experience. By provoking a stop, trying a button, or observing a response, users begin to translate abstract capability into situated understanding. In this sense, tinkering is not only a matter of curiosity. It is also a way of making the robot answerable to the user's own body, timing, and judgment. That is why these acts can be so important for later trust and use. They help transform the robot from something merely presented into something learned.

At the same time, these exploratory acts are not without ambiguity. As the material also shows, actions that begin as curiosity can still interrupt workflow, create inefficiencies, or appear as mockery from another perspective. Guests taking plates from the wrong shelves, rearranging items to see how the robot responds, or stepping repeatedly into its path may be trying to understand the system's logic, yet they also interfere with its operation (Report II, p. 17). This ambiguity is important, but it does not weaken the broader point. On the contrary, it shows that learning the system is itself a negotiated process. People do not always learn robots in the ways designers intended. They learn them through situated interaction, and those interactions can be both productive and disruptive at the same time.

Taken together, these cases suggest that tinkering should not be understood simply as misuse or a deviation from normal operation. It is one of the practical mechanisms through which robotic systems become intelligible to their users. People learn robots bodily, experimentally, and incrementally. They test safety, probe responsiveness, and try to build up enough familiarity to know what the system is and how to act around it. This is why informal skill-building belongs

at the start of the chapter's empirical analysis. Before sabotage becomes resistance, and before interventions become workarounds, there is often a prior moment in which people are simply trying to learn the system by engaging it directly.

2. Making the system workable: workaround, correction, and risky compensation

If the previous section showed how people learn robotic systems by testing and probing them, the cases gathered here show a slightly different form of engagement: what happens when people no longer merely explore the system, but begin to adjust, compensate for, and quietly repair it in order to make it workable in practice. Across our material, this is one of the clearest signs that robotic systems rarely arrive as complete solutions. They enter workplaces and public settings with routes, interfaces, hierarchies, and safety assumptions already built in, yet these rarely fit everyday reality perfectly. People therefore intervene not only to understand the technology, but to keep work moving, restore flow, and align the robot with conditions that are more variable and demanding than formal design tends to anticipate. Many acts labelled as sabotage or tinkering are therefore better understood first as practical attempts to make rigid systems work under real conditions.

This pattern is already visible in the chapter's own discussion of "unintended use." As Rosa explained in the safety-related material, some uses fall outside intended design without necessarily becoming critical failures (Report III, p. 26), and the MedCare example of the small kick captures that logic especially well (Report III, p. 51). One employee discovered that when the robot became stuck, giving it a small kick could prompt it to continue moving. This was clearly outside intended procedure, yet it was not presented as malicious. It emerged through practical experience as a local solution to a local problem. Analytically, that matters because it shows that users do not always meet robots as fixed systems whose rules either are or are not followed. They often encounter them as technologies that sometimes stall, hesitate, or misalign with workflow, and that therefore invite practical correction. The kick is not best understood first as defiance. It is better understood as a workaround that reveals a gap between formal system logic and what is needed to keep work going.

The same point appears in a more distributed way at HandsON. Fiona described a hierarchy for handling the AGVs: the technicians in "facility" repaired and updated them, a rotating responsible person on the warehouse floor handled first-line intervention, and ordinary employees were expected simply to follow traffic rules and call the responsible person when an AGV stopped (Report III, p. 41). Yet in practice, employees at the bottom of this hierarchy often skipped the formal procedure and tried to help a stopped AGV themselves. Fiona's remark is telling: some wanted "a bit more responsibility," but sometimes, if they had just left the robot standing, it might have sorted itself out (Report III, p. 41). This example is valuable precisely because it sits in the missing middle between harmless exploration and resistance. The employees are not primarily trying to undermine the AGV. Nor are they merely playing with it. They are trying to restore workflow by stepping into a gap between the system's formal repair logic and the pressure of practical work. Their intervention may help, or it may do more harm than good, but in either case it is best understood first as compensatory action.

A similar ambiguity appears in the restaurant material. Emma explained that guests would sometimes take plates from the wrong shelves or rearrange items to see how the robot would respond (Report II, 17). On one reading, this is simple

curiosity: people are testing the system in order to understand its logic. On another, it directly interferes with the robot's assigned task and disrupts its workflow. But before deciding whether such actions are exploratory, rude, or obstructive, it is analytically more useful to notice what they reveal. They reveal that the system's practical logic is not self-evident to those around it. Guests do not simply recognise the robot's task order and respect it automatically; they probe it, challenge it, and, in doing so, expose the fact that the robot's working assumptions are still fragile within the social setting it inhabits. Rearranging plates is therefore not only an interruption. It is also a small experiment that makes visible the incomplete fit between formal design and lived use.

The MedCare "tricking" case sharpens this point even further because it shows how such interventions can quickly move from practical engagement to real risk. Andy described an employee who somehow "tricked" the robot into believing it had or had not already placed a container in a location, after which it tried to push another container into the same place and nearly knocked the trolley right off (Report III, pp. 52-53). What matters analytically is that the intention remains unclear. It is not obvious whether this was sabotage, tinkering, a mistake, or an improvised attempt to manage the system. Yet that uncertainty is precisely why the example belongs here. It shows that many interventions arise because humans are continuously negotiating with robotic systems whose internal assumptions and practical effects are not fully transparent. The act may have been exploratory or careless rather than malicious, but it nevertheless had material consequences. This is why the middle ground of workaround and compensation is so important: it is where practical attempts to make systems workable can also become risky.

Seen across the reports, these examples are not isolated. The chapter material already notes that technicians and operators experiment with route settings, speed limits, no-go zones, layouts, storage positions, and handover routines until robots fit existing workflows, and that workers often create unofficial tools such as spreadsheets, scripts, or labels to supplement vendor interfaces and MES systems. At FarmGO, engineers and farmers test sensor configurations, hazard zones, and operating modes under mud, crop, and wildlife conditions (Report IV, pp. 7-8); at ShipGO, captains run SPOS routes alongside their own judgment to decide when the system is helpful and when it is not (Report IV, p. 22). These are all instances of the same broader pattern: people do not simply adopt robotic systems as given. They complete them, correct them, and quietly reshape them so they can function under real conditions. In this sense, workaround is not peripheral to automation. It is part of how automation is practically sustained.

This is also the point at which the section ties back most clearly to Sim2Real. The Sim2Real chapter argued that human actors continuously sustain alignment between model and world by noticing drift, adjusting maps, compensating for wear, and feeding local experience back into formal representation. The same logic appears here in a more interactional

and work-practical form. Users are not only learning the system; they are compensating for the mismatch between the system's built-in assumptions and the realities of practice. A kick that gets a stuck robot moving, an unauthorised attempt to help an AGV, or the quiet adjustment of interface routines are all small acts of model–world repair. They are ways of making a formal system workable in a lived environment that exceeds what was stabilised at design time. This is why these interventions should not be treated too quickly as either sabotage or tinkering. First and foremost, they are signs that autonomy in practice remains dependent on local human adjustment.

At the same time, the material shows why this compensatory work is inherently ambivalent. Some interventions clearly improve safety, flow, and usability by bridging the gap between rigid system logic and messy real conditions. Others begin to normalise risky habits, bypass formal safeguards, or conceal deeper problems in the system. As the chapter notes, a worker who repeatedly overrides an AGV stoppage may be compensating for an over-conservative setting but may also be quietly normalising behaviour that bypasses real hazards (Report III, p. 56). From a cross-case perspective, workaround is therefore neither simply good nor bad. It is better understood as a response to misalignment that may be wise, necessary, risky, or all three at once.

What this middle ground makes visible is that many human interventions begin not in hostility toward robots, but in the practical effort to keep them useful. People work around robotic systems because the systems do not yet fully fit the worlds they have entered. But once such interventions become repeated, necessary, or frustrating, they also open onto a different question: when does making the system workable turn into a sign that the system itself is not fully accepted, trusted, or legitimate? It is at that point that the chapter moves from workaround toward resistance.

3. Resistance, refusal, and legitimacy crises

We now move from workaround toward resistance. As long as people adjust, compensate for, and quietly repair robotic systems in order to keep them usable, their interventions still remain broadly aligned with the system's practical continuation. They may bend rules, bypass procedures, or develop unofficial routines, but they do so in order to make the technology function within everyday work. Resistance begins when that alignment weakens. Here interventions are no longer aimed primarily at making the system workable, but at limiting it, sidelining it, contesting it, or demonstrating that it does not deserve the place it has been given. This is where sabotage becomes analytically distinct from tinkering: not because the outward action is always different, but because the relation to the system changes. The robot is no longer treated mainly as something to be learned or helped along, but as something experienced as threatening, imposed, or misaligned with people's own sense of good work, responsibility, and control.

A central pattern across the material is that sabotage appears most clearly where workers experience a loss of autonomy. As the empirical material shows: sabotage emerges in situations where the system is felt to reorganise work, redistribute responsibility, or challenge existing forms of judgment and professional identity. This is especially visible in the agricultural literature from Martin et al. (2022), where they demonstrate that robots change both work organization and the meaning of work. With automated milking systems in the agricultural field, the physical workload is reduced, but new tasks emerge, such as monitoring alarms and managing data. According to their research, this emergence of new work tasks can lead to "technostress" and increased mental strain (Martin et al. 2022). When employees (farmers

in this case) experience that work becomes more surveillance-oriented and less craft-based, it can create a sense of loss of identity and autonomy. In such a context, sabotage can be seen as a protest against the changed conditions. In such a setting, sabotage can be understood not simply as irrational opposition to technology, but as a reaction to altered work conditions experienced as imposed from outside. If automation changes the meaning of work, shifts control, or turns skilled practitioners into passive monitors of technical systems, then resistance becomes intelligible as a response to that transformation.

The MedCare case provides one of the clearest empirical examples of this process. Andy explained that the first AGVs introduced there were not only technically limited but also received with considerable annoyance because employees feared they would be replaced (Report III, pp. 52-53). Workers quickly discovered that the robots would stop if something was placed in front of them, and they began using this weakness to obstruct their operation. The result was that the AGVs no longer appeared efficient or productive, and the project was eventually scrapped. This is an important case because it shows that sabotage does not emerge only when systems malfunction in technical terms. It can also emerge when people experience the technology itself as a challenge to their place in the organisation. In such cases, sabotage becomes a way of contesting the future the robot seems to represent.

A related dynamic appears in Nick's examples from ElectroFlight. He described how workers sometimes put tape in front of sensors, used stickers to interfere with them, or even created special tools to gain access inside fenced robot areas (Report III, pp. 14-15). On one level, these acts can be read as violations of formal safety arrangements. But the material also suggests that they are more than that. Nick associated some of them with dissatisfaction and concerns about job loss, while elsewhere he also noted that similar acts could arise from workers feeling overly confident in their own control over the system (Report III, p. 15). This is precisely why the move from workaround to resistance is analytically important. The same kinds of actions—bypassing sensors, entering restricted zones, overriding safety procedures—can shift from practical compensation into resistance when they become ways of asserting human control against a system experienced as restrictive, frustrating, or socially threatening. Resistance is therefore not defined only by the form of the act, but by the broader situation in which the act is embedded.

This example from ElectroFlight highlights how ambiguous acts of sabotage and tinkering can be. Because we did not speak directly with the employee who interfered with the safety beam sensors, we can only speculate about their intentions, and whether the act was deliberate sabotage or simply an instance of tinkering. However, determining the label is not the central issue. What matters is that the action is diagnostic: it signals that something significant is happening and deserves attention. As Nick suggests, the employee may have developed a sense of trust in the robot, feeling comfortable working alongside it while it was operating (Report III, p. 15). From this perspective, the act can be understood as tinkering, with the employee adjusting the safety system to better fit the workflow. Nevertheless, the employee is still interfering with a critical piece of safety equipment. If a coworker were to approach the robot unaware that the sensor had been blocked, they could reasonably expect the system to stop and might therefore be put at risk. In this way, an action intended as tinkering can instead be perceived as sabotage. The purpose of this example is not to assign blame or impose a strict categorization, but to illustrate how such interactions with robotic systems are sites of

learning. They function as diagnostic moments: opportunities to identify underlying assumptions, practices, and risks, and to learn from them.

The maritime cases sharpen this argument further because they show that resistance does not always take an active or destructive form. Frank's refusal is a good example. He said that if autonomous ships were introduced, many captains would not want to work on them because they would not want to bear responsibility if something went wrong (Report IV, p. 19). This is significant because it reveals a more passive form of sabotage or refusal: not attacking the system directly but withholding cooperation from it. Here resistance arises from the misalignment between technological autonomy and established structures of accountability. Captains are not simply sceptical because they dislike innovation. They are sceptical because the technology threatens to redistribute control without clearly redistributing responsibility. In this sense, refusal is also a form of resistance to autonomy, and one that shows how deeply sabotage and tinkering are tied to the themes of trust and accountability developed in the previous chapters.

This broader logic is also reflected in the spectrum section and conclusion already present in the chapter. There, resistance is described as including blocking robots, parking them in corners, unplugging them, "forgetting" to use them, or simply reverting to manual work. In public settings, it may even take symbolic forms, such as a customer flipping off the robot at SHOPFLOOR (Report II, p. 24). These acts may not always disable the system entirely, but they are significant because they signal that the robot has not achieved legitimacy in the setting it has entered. When people actively undermine a system, it is often because they experience it as unfair, threatening, pointless, or misaligned with their own values and responsibilities. This is a much stronger formulation than treating sabotage merely as negative behaviour. It makes sabotage analytically useful, because it shows where the social and organisational basis of robotic adoption begins to crack.

Seen from this perspective, resistance is not outside the story of autonomy. It is part of how autonomy is negotiated in the wild. Robots do not simply enter neutral spaces where they are either accepted or rejected based on technical merit alone. They enter workplaces, professions, and public settings already structured by norms of competence, authority, identity, and responsibility. Where robotic systems fit poorly with these arrangements, people may first tinker, then workaround, and finally resist. The move from workaround to resistance therefore marks a shift in the practical relation between humans and the system: from trying to make the technology workable to questioning whether it should be made workable on those terms at all.

At the same time, the chapter's material also suggests that such resistance is not fixed. As the conclusion notes, sabotage and tinkering often have a temporal dimension. They may be especially prevalent in early phases of implementation and diminish as familiarity, trust, and routines develop. In several settings, including MedCare, relatively seamless collaboration later emerged even within quite complex workflows (Report III). This matters because it prevents the section from becoming too static or moralised. Resistance does not simply reveal anti-technological attitudes. It often marks an unstable phase in which legitimacy, fit, and responsibility are still being negotiated. But precisely for that reason it remains analytically valuable. It tells us where adoption is fragile, where design has failed to secure acceptance, and where the introduction of autonomy is experienced less as support than as disruption.

Taken together, these cases suggest that sabotage should not be treated merely as malicious misuse. However the empirical examples of sabotage shows that something that there is something in the human-robot interaction (HRI) that is not working and should be further understood. Importantly, this does not primarily point to shortcomings in the robot itself or its manufacturer, but to challenges in how robotic systems are introduced, integrated, and managed within organisations. People may resist when systems are experienced as threatening their autonomy, distorting established work practices, redistributing responsibility in unclear or unfair ways, or demanding forms of trust that have not been sufficiently established. In this sense, sabotage often reveals less about hostile users than about misalignments between the technology, the work environment, and organisational decisions surrounding its deployment.

This highlights the importance of implementation: organisations adopting robotic systems play a key role in shaping how work changes, how responsibilities are communicated, and how trust is built among employees. From this perspective, acts that might be labelled as sabotage can be understood as responses to these broader conditions. For this reason, sabotage should not be treated as an isolated or marginal phenomenon, but as one end of the spectrum of interactions that also includes playful testing, adaptation, and informal learning. Seen in this way, such actions become diagnostic, not of individual intent alone, but of how well the overall socio-technical integration is functioning.

Discussion of Sabotage/Tinkering in Self-Adaptive Robotic Systems

From our empirical research it has become clear that sabotage and tinkering are not only external disturbances or unintended use, but they also show integral forms of engagement. Across the fieldwork, people do not relate to robots in binary terms of acceptance or rejection. Instead, they engage with them in ways that range from exploratory interaction and practical adjustment to resistance and, in some cases, deliberate disruption. These forms of engagement become particularly significant in adaptive systems, where behaviour is expected to evolve over time in response to data, feedback, and environmental input.

Tinkering represents one of the most immediate and pervasive ways in which users interact with such systems. It often takes the form of exploratory, curious, and hands-on engagement through which people test system boundaries, probe responses, and develop situated understanding. This kind of interaction is not primarily oppositional. Rather, it reflects an attempt to make opaque or complex systems intelligible in practice. In environments where robots arrive as partially understood technologies, tinkering becomes a necessary condition for learning. Users build familiarity not through manuals or formal training alone, but through interaction, by observing how the system reacts, where its limits lie, and how it fits into existing workflows. In this sense, tinkering is closely tied to adaptation, not only at the level of the system but also at the level of the user.

However, in many cases, this engagement moves beyond exploration into what can be described as compensatory intervention. Here, users actively adjust, support, or work around the system in order to make it function under real-world conditions. These actions may include restarting systems, bypassing overly sensitive safety features, repositioning robots, or developing informal routines that align system behaviour with practical demands. While such interventions

may technically fall outside intended use, they are often essential for maintaining operational continuity. In the context of self-adaptive systems, this might raise an important question: much of the system's apparent functionality depends not only on its internal adaptive mechanisms, but on continuous human correction and support. Adaptation, in practice, may therefore be less a purely "self"-driven process than something that emerges through the ongoing interplay between human intervention and system behaviour.

At the same time, not all engagement remains supportive. As systems are introduced into complex social and organisational environments, forms of resistance may emerge. These can include subtle acts such as ignoring the system, reverting to manual practices, or informally deprioritising its use, as well as more visible actions like blocking, interrupting, or mocking the robot. Such behaviours do not necessarily aim to destroy the system, but they signal a weakening alignment between the system's design and the users' expectations, values, or working conditions. In adaptive systems, this misalignment can be particularly pronounced if system behaviour changes over time in ways that are difficult for users to anticipate or understand. What is framed technically as "adaptation" may, from a user perspective, appear as unpredictability or loss of control.

In more extreme cases, engagement takes the form of deliberate sabotage, where users actively seek to disrupt or undermine the system's operation. While this may appear as irrational or malicious behaviour from a technical standpoint, the empirical material suggests that such actions often point to deeper issues of legitimacy. Sabotage tends to arise where systems are experienced as imposed, misaligned with professional practices, or threatening to workers' autonomy, identity, or sense of responsibility. In this sense, sabotage is not only an act against the system, but also a signal about the conditions under which the system has been introduced and is expected to operate.

For self-adaptive robotics, these forms of engagement could raise a set of critical questions. If systems are designed to learn from interaction, then it might become important to consider which kinds of interaction are actually recognised as meaningful input. Much of the adjustment, correction, and improvisation carried out by users remains invisible to the system, even though it plays a central role in making the system workable. At the same time, actions such as resistance or sabotage may be registered only as noise or error, rather than as indications of deeper misalignment. This suggests that adaptive capacity might not only be a matter of improving algorithms or increasing data flows, but also of recognising how human practices contribute to, challenge, or reshape system behaviour.

Ultimately, sabotage and tinkering highlight that self-adaptive systems do not operate in isolation from the social worlds into which they are deployed. Adaptation is not confined to internal system processes, but unfolds through ongoing interaction with users who interpret, adjust, and sometimes contest the system. These engagements make visible the limits of purely technical understandings of adaptation and point instead to a more relational view, in which system behaviour is continuously shaped by the interplay between computational logic and situated human practice.

Conclusion on Sabotage/Tinkering

Taken together, this chapter shows that sabotage and tinkering should not be understood as marginal disturbances at the edge of robot use. They are central ways in which robotic systems are encountered, negotiated, and made meaningful in practice. Across the cases, people do not simply accept or reject robots as finished technologies. They test them, learn them, help them along, work around them, resist them, and sometimes deliberately undermine them. For that reason, sabotage and tinkering are best understood not as fixed and separate categories, but as points along a spectrum of engagement with robotic systems. What matters is not only the visible action, but the practical relation to the system that the action expresses.

At one end of this spectrum lies tinkering as exploratory engagement. The chapter has shown that people often learn robotic systems by probing them: stepping in front of them, pressing buttons, testing safety responses, or experimenting with unfamiliar interfaces in order to understand how they behave. In this sense, tinkering is not a failure of adoption, but part of adoption itself. People rarely encounter robots as transparent and self-explanatory tools. They make them intelligible through bodily testing, informal experimentation, and situated learning. This is one reason why systems that leave no safe room for exploration are often experienced as more fragile or more alienating. Tinkering helps users develop the practical familiarity through which robotic systems can begin to become workable at all.

Moving further along the spectrum, the chapter has shown that many interventions are not merely exploratory, but compensatory. People do not only test robots; they also adjust them, help them along, and quietly repair them in order to make them function under real conditions. In workplaces especially, users develop local hacks, workarounds, and informal correction practices that bridge the gap between formal system logic and the contingencies of actual work. These interventions may sit uneasily with official procedures, but they are often part of what keeps robotic systems operational in practice. Tinkering in this sense becomes a form of situated design. Users do not simply adopt what engineers have built; they complete it, reshape it, and align it with the realities of workflow, safety, and everyday use. At the same time, these practices are ambivalent. Some improve safety and robustness; others normalise risky habits or hide deeper problems in the system. Sabotage and tinkering therefore cannot be labelled simply good or bad. They are symptoms of deeper misalignments and mechanisms through which systems are adapted, sometimes wisely and sometimes riskily.

At the other end of the spectrum lies resistance and sabotage. Here interventions are no longer primarily aimed at learning the system or making it workable, but at limiting it, contesting it, or demonstrating that it does not deserve the place it has been given. The chapter has shown that such sabotage often appears where workers experience robotic systems as unfair, threatening, pointless, or misaligned with their own values and responsibilities. It may take active forms, such as blocking sensors or obstructing robots, but it may also take quieter forms of refusal, disregard, or symbolic rejection. In this sense, sabotage is analytically important not because it is always destructive, but because it often signals a legitimacy crisis. When people actively undermine a system, something important is usually at stake: a gap between promised and actual performance, between formal procedures and real work, or between technical control and workers' sense of responsibility, autonomy, and identity.

Viewing Sabotage and Tinkering as a spectrum shows that outwardly similar actions can have very different meanings depending on context and intention. Stepping in front of a robot may be exploratory, corrective, mocking, or obstructive. Overriding a stoppage may be a sensible response to an overly conservative system or the beginning of a dangerous normalisation of risk. Intention matters, but intention is not always directly observable. This is why sabotage and tinkering might best be treated as a relational position within a broader field of engagement, rather than as fixed behavioural types. The distinction therefore does not simply classify user conduct. It reveals how robotics in practice is shaped through ongoing negotiation between technical systems and human actors who are trying to understand, adapt, and sometimes resist them.

At the same time, the chapter has shown that these behaviours are not static. In several cases, sabotage and tinkering appeared most strongly in early phases of implementation, when systems were still unfamiliar, mistrusted, or poorly integrated into local routines. Over time, people often developed new rhythms and working relations with the robots, and the need for overt resistance or constant informal intervention could diminish. This temporal dimension is important because it prevents sabotage and tinkering from being reduced to simple anti-technology attitudes. They often indicate unsettled phases of introduction in which legitimacy, trust, and fit are still being worked out. Yet precisely for that reason they remain analytically valuable. They show where the social and organisational conditions of adoption are fragile, and where formal design has not yet secured practical acceptance.

For the larger argument of this report, sabotage and tinkering therefore do important conceptual work. They show that robot autonomy is never just a matter of how much control has been moved into the machine. It is also about how much room humans have to question, reshape, and, when necessary, resist the systems that enter their workplaces, ships, farms, hospitals, and public spaces. Autonomy in the wild is not simply enacted by robots. It is co-produced through the human practices that learn, support, destabilise, and contest them. Sabotage and tinkering make that especially visible because they mark the points where technical ambition meets lived reality most directly.

Taken together, these themes suggest that sabotage and tinkering are diagnostic categories. They show where robotic systems remain opaque, where formal design does not fully match practical reality, and where the introduction of automation becomes socially contested. This also connects the chapter directly to the previous sections of the report by showing what people actually do when tensions become concrete in everyday interaction: they test, adapt, bypass, correct, resist, and sometimes deliberately disrupt the system. In that sense, sabotage and tinkering are not outside autonomy. They are among the practices through which autonomy is negotiated in the wild.

Use, Implementation, and Design Challenges

While Report V is primarily concerned with the complex interactions that arise in real-world environments through the study of autonomous robots “in the wild,” the RoboSAPIENS project ultimately aims to deliver commercially viable software tools. These tools, designed to enable open-ended software adaptation in robotic systems, are intended for deployment in precisely such “wild” environments. Crucially, they must not only work in these environments but also operate in ways that remain safe, efficient, and trustworthy. Achieving this requires recognising that the primary users of the project’s outputs are not necessarily the end users examined in this report. Rather, at least initially, they are more

likely to be the organisations and companies responsible for the development, integration, implementation, and operation of these systems.

The immediate users of the RoboSAPIENS project tools are therefore engineers, integrators, and software specialists working within industry-specific consultancies or robot manufacturing firms. The ultimate subjects of their impact, however, are more diffuse and heterogeneous groups, such as those represented by our informants. Examples cited in RoboSAPIENS project literature include production workers, maritime crews, maintenance personnel, and operators across a range of industries. Like those studied here, these groups are diverse and embedded within distinct organisational, cultural, practical, and environmental contexts. As such, there is no single, stable definition of the "end user" that can be readily translated into system requirements for those responsible for implementing RoboSAPIENS tools.

At the same time, although self-adaptive robotics must contend with a shifting landscape of users whose needs and practices evolve over time, deployment will typically occur within specific user groups and contexts, each with situated expectations of what counts as "safe," "trustworthy," and effective. This creates a significant challenge for projects applying MAPLE-K architectures as their tools must not only be usable by highly specialised experts but also made meaningful and applicable to practitioners operating across diverse organisational settings.

More specifically, the RoboSAPIENS project literature (Larsen et al., 2024) describes the development of software tools intended to support the adaptation of Cyber-Physical Systems (CPSs), (Larsen et al., 2016), in which CPS engineers are equipped with integrated, model-based toolchains to support both design and deployment. As such, these tools occupy a critical position between system design and system use as they shape how robotic systems are configured, evaluated, and adapted during development and implementation, prior to their eventual encounter with end users in real-world environments.

In addition to the interest group represented by the International Federation of Robotics, the RoboSAPIENS project proposal explicitly identifies end users as manufacturing firms engaged in risk assessment, remanufacturing companies, maritime operators, and organisations responsible for managing robot fleets. These categories align with the end-user groups represented across the project's four user cases, with the broader aim of reducing the workload of human workers, including production staff, maritime personnel, and manufacturing floor operators.

Integrating MAPLE-K project architectures and tools into a variety of robot control systems raises further questions. While this integration involves technical practices grounded in complex software and robotic engineering, this also involves the kinds of lived, relational, and situated practices examined in this report. These processes are, moreover, shaped and constrained by organisational, market-related, and economic factors, where for instance, small and medium-sized enterprises may face limitations in terms of resources and technical capacity. Beyond these considerations, the challenge of adapting robot behaviour in uncertain environments introduces additional concerns around accountability, including issues of verification, validation, and predictability.

In general, the findings of this report emphasise that aspects such as safety and trust are not simply embedded technical properties but depend on relational dynamics of human engagement and environmental conditions. Closely aligned with this are perspectives on technological adoption, in which the uptake of new technologies such as MAPLE-K is shaped by organisational readiness, training, and accountability structures. In this sense, the reasons for not ultimately achieving open-ended, trustworthy self-adaptation may be less technical than social in character. A lack of in-house expertise, for example or difficulties integrating new paradigms into existing workflows, or uncertainty about return on investment. The introduction of adaptive systems therefore requires not only new tools, but also new competencies, practices, and forms of coordination.

Whilst beyond the immediate scope of this report, considering MAPLE-K project as an innovation in the field of robotics also raises questions of market and innovation readiness. These include, for example, the challenge of recouping high development and implementation costs, as well as the broader difficulties of scaling and commercialising such systems. Adoption is also contingent on whether the RoboSAPIENS project can demonstrate clear value. In the project's own terms, this involves achieving self-adaptation to address environmental and human uncertainty while maintaining guarantees of safety, efficiency, and trustworthiness. However, these benefits must also be made visible and credible to decision-makers as evidence of improved performance and user satisfaction. This has direct implications for the RoboSAPIENS project in the long run as its advantages must not only function effectively but also position themselves within existing market structures. This includes demonstrating compatibility with current industry practices alongside supporting integration into existing toolchains, and aligning with the expectations of stakeholders such as those associated with the International Federation of Robotics.

Another set of challenges, more closely aligned with the findings of this report, concerns regulation, ethics, and public trust. As systems become increasingly autonomous, requirements for safety assurance, accountability, and regulatory compliance also increase significantly. Work by M. Bilal (Bilal et al., 2025) shows that regulatory uncertainty and the absence of clear legal frameworks can slow adoption, while ethical concerns, such as workforce displacement and data privacy, shape how these technologies are perceived and accepted. These issues resonate strongly with this report, particularly its emphasis on trust and the enacted dimensions of safety. They also reinforce the importance of the RoboSAPIENS project's focus on trustworthiness. However, the findings of the report would suggest that trustworthiness cannot be engineered through technical systems alone but needs to be equally established through aspects such as transparency and human expectation.

This suggests the need to combine technological design with contextual insights of use, such as those highlighted in this report. A range of well-established user-centred design approaches provide methodologies capable of informing complex design problems such as those addressed by the RoboSAPIENS project. Through these approaches, empirically grounded insights, such as how safety is enacted in practice, can inform the design and evaluation of software tools. User-centred knowledge could therefore become an integral basis for design processes and decision-making. In this sense, what counts as "safe" or "trustworthy," as understood through the experiences, expectations, and behaviours of those who engage with robotic systems, can offer important guidance for how such systems are engineered and

developed. User and social orientated design approaches unearth the complex relationships between design and use. Software architectures and tools are not to be understood as neutral instruments. They actively shape how robotic systems are configured and how assumptions about safety, robustness, and trustworthiness are operationalised. The Social Construction of Technology (SCOT) (e.g. Bijker 2015) perspective for example, suggests that technologies achieve stability and market viability not through technical optimisation alone, but through their changing alignment with the meanings and practices of relevant social groups. In this light, the project's commercial implications such as market position and competitiveness depend on whether MAPLE-K software tools can be understood and adapted by different stakeholders. This, in turn, requires sensitivity to the different contingencies of real environments, rather than reliance on abstract or idealised models of use.

This report suggests that robust and safe systems are ultimately those that can be tested against the unpredictability and situated practices of everyday work. The RoboSAPIENS project's intended use of digital twin technologies within the Legitimate phase offers a mechanism for simulating and evaluating system behaviour. However, a key question remains: to what extent can such simulations capture the richness of actual human environments? Bridging this Sim2Real gap requires the ongoing incorporation of empirical insights into the design and testing of validation processes. Insights from the field of Participatory Design, for example, highlight the importance of engaging with the complexity of real-world contexts through the involvement of multiple stakeholders.

Ultimately, the impact of the RoboSAPIENS project, whether practical or commercial, depends on its ability to reconcile the promise of self-adaptive systems with the complexities of uncertain environments and human interaction. Its tools must be sufficiently robust to support reliable system behaviour, yet sufficiently flexible to accommodate the diversity of real-world contexts. By grounding development and evaluation in the contextual insights presented in this report, and by engaging with user-centred design perspectives, the project may be better positioned to advance self-adaptive robotics through the design of systems that are meaningfully safe, trustworthy, and adoptable in practice.

Conclusion Report V

This report has argued that robot autonomy should be considered not as independent from humans, but as a distributed and negotiated accomplishment. Across the *Robots in the Wild* series, robots have been followed not only in idealised laboratory settings, but also in the everyday environments in which they are expected to function: conferences, restaurants, museums, warehouses, hospitals, farms, and maritime contexts. Viewed across these settings, autonomy does not emerge as a stable technical threshold that, once reached, allows the machine to operate independently. Rather, it appears instead as something that is continuously achieved, supported, interrupted, recalibrated, and at times undone in practice. Robots do not operate alone in any simple sense. Their functioning depends on human oversight, organisational routines, environmental conditions, infrastructures, standards, and situated judgement. And as noted one of the strongest findings across the cases is that increasing robot autonomy does not remove responsibility; it redistributes it. In this sense, autonomy in the wild is not stand-alone autonomy, but distributed autonomy.

The five themes developed in this report show why autonomy must be understood in this broader way. The chapter on safety demonstrated that safety is not only a design-time property to be built into systems, but something that must remain legible, accountable, and revisable in deployment. The chapter on trust showed that people do not only ask whether systems work, but whether they can rely on them physically, functionally, and informationally in practice. The chapter on anomalies showed that deviations are not rare faults at the margins of otherwise smooth automation, but recurring moments where the practical limits of autonomy become visible and where human intervention becomes necessary. The chapter on Sim2Real showed why this continues to happen: robots do not encounter the world directly, but through maps, models, and simulations that are always partial and therefore always in need of practical repair. Finally, the chapter on sabotage and tinkering showed that human interventions are not merely interference from outside the system, but among the ways robotic systems are actually learned, completed, contested, and made workable in everyday life.

Taken together, these findings suggest a more substantial overall claim. Robot autonomy is not simply reduced or made less adequate by human involvement; it is indeed co-constituted through it. Human actors are not external obstacles or passive elements in autonomous systems. They are part of the socio-technical arrangements making robotic action possible, safe, interpretable, and useful. This does not suggest that autonomy is an illusion, nor that technical advances in sensing, planning, simulation, and adaptive control are unimportant. It means rather that autonomy, once deployed in real settings, always becomes relational and distributed. It depends not only on what the robot can monitor, analyse, plan, legitimate, execute, and remember internally, but also on whether humans can understand what it is doing, decide when to trust it, correct it when it fails, and integrate it into the practical and moral orders of everyday work.

This is also where the report's contribution to RoboSAPIENS becomes readily apparent. The project's ambition is not modest: to develop robots capable of open-ended self-adaptation under unknown change while remaining safe, robust, and trustworthy. The ethnographic material presented here does not rule out that ambition. It shows that trustworthy self-adaptation cannot be understood only as a matter of improving internal adaptation loops, runtime verification, or trustworthiness checking, however crucial those are. If adaptive autonomy is to work outside controlled settings, it must also be designed and configured in ways that remain legible, accountable, and workable for the people who encounter it. The challenge is therefore not only how robots adapt, but how adaptation itself becomes understandable, acceptable, and open to intervention in practice.

From this perspective, the report points toward a revised understanding of successful self-adaptive robotics. Success cannot only be measured by whether a system technically adapts itself to changing conditions alone. It must also be measured by whether that adaptation can coexist with human judgment, organisational responsibility, and the practical variability of lived environments. A self-adaptive system that changes behaviour in ways users cannot interpret may be formally adaptive, yet socially brittle. A system that updates its models but leaves people unsure when to trust or override it may be computationally impressive, yet practically unworkable. Conversely, systems that allow situated knowledge to feed back into design, and reconfiguration that make anomalies visible, that support meaningful human

oversight, and that remain accountable under changing conditions are more likely to become not only adaptive, but adoptable.

This also clarifies the significance of the report's anthropological and STS-informed approach. The conceptual themes developed across the chapters are not alternatives to engineering knowledge, but extensions of it. They make visible aspects of robotic operation that are often backgrounded when autonomy is treated only as an internal system capacity: how risk is negotiated, how trust is calibrated, how anomalies are interpreted, how simulations depend on human compensation, and how interventions may express learning, correction, or resistance. In doing so, they provide a vocabulary for understanding why autonomous systems so often behave differently in the wild than in demos, simulations, or design documents. They also suggest that these differences are not merely unfortunate implementation problems. They are part of what autonomy becomes when it enters the world.

The broader implication is therefore that future work on self-adaptive robotics should not aim simply at removing humans from the loop. The more productive ambition in future research is rather to redesign the loop itself. Rather than imagining progress as a steady reduction of human involvement, RoboSAPIENS and related efforts might better ask how adaptive systems can support better forms of involvement: forms of involvement that are not ad hoc compensation for technical fragility, but structured routes for interpretation, oversight, correction, and learning. This would mean treating human practice not as noise around the system, but as part of the environment to which robust autonomy must remain responsive. It would also mean taking seriously that some forms of local intervention, workaround, and even resistance are not signs of failure alone, but sources of design knowledge about where systems remain unfinished or misaligned with practice.

Report V therefore closes with a simple but consequential conclusion: the future of autonomous robotics does not lie in achieving independence from humans, but in developing forms of adaptive autonomy that remain safe, trustworthy, intelligible, and workable within human worlds. Autonomy in the wild is not the disappearance of human judgment, but its redistribution across new technical and organisational forms. If RoboSAPIENS is to realise its ambition of trustworthy self-adaptation, then that ambition must be pursued not only through better adaptive architectures, but through a deeper engagement with the practical conditions under which autonomy is lived, negotiated, and sustained. The central lesson of this report is thus not that autonomous robots cannot work. It is that they work differently than visions of autonomy often suggest, that is through distributed relations, ongoing adjustment, and the continuous co-production of machine capability and human practice.

References

Akrich, Madeleine (1992). The De-Scription of Technical Objects. In Wiebe E. Bijker and John Law (eds.), Shaping Technology / Building Society: Studies in Sociotechnical Change. Cambridge, MA: MIT Press.

Asimov, Isaac. (1942/1950). *Runaround. In I. Asimov (Ed.), I, Robot.Garden City:Doubleday.*

Barricelli, B. R., & Fogli, D. (2024). Digital twins in human-computer interaction: A systematic review. International Journal of Human–Computer Interaction, 40(2), 79-97.

Bijker, Wiebe E. (2015). "Technology, Social Construction of". In Wright, James D. (ed.). *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Elsevier. pp. 135–140. doi:10.1016/B978-0-08-097086-8.85038-2. ISBN 978-0-08-097087-5.

Bilal, M., Ameer, S., Awan, A. S., Mumtaz, A., Gul, S., & Shahbaz, N. (2025). Regulating the Future: A Holistic Model for Safe and Ethical Deployment of Robotic Systems in Emerging Industries. *Journal of Computing & Biomedical Informatics*, *9*(01).

Birch, A. R., Kristensen, M. H., Wright, T., Bollmann, Y., Scharping, R., Morrison, A. L., Hasse, C., & Esterle, L. (2026, February 27). *Trust and anomalies in the design of safe robotic self-adaptation* (Version 2) [Preprint]. TechRxiv. https://doi.org/10.36227/

Perez, Caroline Criado. (2019). *Invisble Women: Data Bias in A World Designed for Men.* Harry N. Abrams (red.)

European Commission. (2019). Ethics guidelines for trustworthy AI. Publications Office of the European Union. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Flyverbom, Mikkel. (2019). Chapter 4: Organizations Gone Transparent. *The Digital Prism.* Cambridge

Görnemann, Otto. (2019). *EN ISO 1200 and its relation to the Machine Directive.* SICK AG.

Hasse, C. (2015). *An anthropology of learning. On Nested Frictions in Cultural Ecologies.* Dordrecht.

Hasse, C. (2020). *Posthumanist learning. What robots and cyborgs teach us about being ultra-social*. Routledge

Hollnagel, Erik. (2009). *The ETTO Principle: Efficiency-Thoroughness Trade-Off*. CRC Press.

Ihde, Don. 2006. “The Designer Fallacy and Technological Imagination.” In *Defining Technological Literacy: Towards an Epistemological Framework*, ed. John R. Dakers, 121–32. New York: Palgrave Macmillan.

Ingold, T. (2000). *The Perception of the Environment: Essays on Livelihood, Dwelling and Skill.* London: Routledge.

ISO. ISO - Standards . https://www.iso.org/standards.html

ISO 10218:2025 Robotics—Safety requirements for industrial robot applications and robot cells. ISO.

Larsen, P. G., Fitzgerald, J., Woodcock, J., Fritzson, P., Brauer, J., Kleijn, C., ... & Sadovykh, A. (2016, April). Integrated tool chain for model-based design of Cyber-Physical Systems: The INTO-CPS project. In *2016 2nd International Workshop on Modelling, Analysis, and Control of Complex CPS (CPS Data)* (pp. 1-6). IEEE.

Larsen, P. G., Ali, S., Behrens, R., Cavalcanti, A., Gomes, C., Li, G., ... & Zhang, H. (2024). Robotic safe adaptation in unprecedented situations: the robosapiens project. *Research Directions: Cyber-Physical Systems*, *2*, e4.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Leenes, Ronald & Federica Lucivero. (2015). Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behvaiour by Design. *Law, Innovation and Technology.*

Martin, T., Rives, J., Cavet, C., Moraine, M., Bohanec, M., Debaine, F., … & Hostiou, N. (2022). Robots and transformations of work in farm: A systematic review of the literature and a research agenda. *Agronomy for Sustainable Development, 42,* 96.

Massaguer Gómez, G. (2025). Should we trust social robots? Trust without trustworthiness in human-robot interaction. *Philosophy & Technology*, *38*(1), 24.

Moustakidis, V., Passalis, N., & Tefas, A. (2024). *A taxonomy of anomalies in AI-powered robotic systems* [Manuscript submitted for publication]. Aristotle University of Thessaloniki.

Nye, David. (2006). Chapter 11: Not Just One Future. *Technology Matters*. The MIT Press.

Official Journal of the European Union (2006) *Machine Directive 2006/42/EC.*

Reissig, L., & Siegrist, M. (2025). From the attitude towards digitalisation in agriculture to the acceptance of future agricultural technologies. *Smart Agricultural Technology, 12,* 101095.

Report I. Robot Autonomy and Conferences. *Robots in the Wild.* RoboSAPIENS. Robots in the Wild – RoboSapiens.

Report II. Robots in Public Space. *Robots in the Wild.* RoboSAPIENS. Robots in the Wild – RoboSapiens.

Report III. Robots in Organisations (Controlled Environments). *Robots in the Wild.* RoboSAPIENS. Robots in the Wild – RoboSapiens.

Report IV. Robots in Dynamic Environments. *Robots in the Wild.* RoboSAPIENS. Robots in the Wild – RoboSapiens.

Roesler, E., Vollmann, M., Manzey, D., & Onnasch, L. (2024). The dynamics of human–robot trust attitude and behavior—exploring the effects of anthropomorphism and type of failure. Computers in Human Behavior, 150, 108008.

Wallace, J., & Hasse, C. (2014). Situating Technological Literacy in the Workplace. I J. Dakers (red.), *New Frontiers in Technological Literacy: Breaking with the Past* Palgrave Macmillan.

Winner, Langdon. (1980). Do Artefacts Have Politics? *Daedalus.*

Robotic safe adaptation in unprecedented situations: the RoboSAPIENS project | Research Directions: Cyber-Physical Systems | Cambridge Core