Grunwald, A. 2025. “Artificial Intelligence: Responsible Innovation in the Face of Potential Gradual Disruptions” [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 68–83.

Armin Grunwald[*]

Artificial Intelligence: Responsible Innovation in the Face of Potential Gradual Disruptions[**]

Submitted: Aug. 13, 2025. **Reviewed:** Oct. 19, 2025. **Accepted:** Oct. 18, 2025.

Abstract: This paper deals with the possibility of gradual disruptions at the societal level in the course of rapidly advancing digitalization and spread of AI. The term “disruption” refers to the sudden breakdown of familiar, previously stable constellations. Expectations of stability, assumptions of continuity, and planning security are shattered, casting the prospects for the future in an uncertain light. The Latin roots of the term mean “bursting,” “breaking,” and “tearing,” semantically referring to the temporal structure of more or less sudden, abrupt events. Seen in this light, the talk of gradual disruption in the title of this article seems conceptually contradictory or paradoxical. However, there are many examples of disruption in the world of technology that were heralded by recognizable but often unnoticed signs, particularly by material fatigue and wear. The daily stress on many technical objects, such as V-belts in older vehicles or bridge structures, gradually leads to wear and degradation. In this sense, the notion of gradual disruption refers to upheavals with significant or even dramatic damage potential that do not occur unexpectedly and suddenly, like a global pandemic or an earthquake, but build up gradually until they finally lead to the disruption of previously stable constellations. I will argue that this type of potential and gradual disruption could emerge in areas of digitalization and AI. Examples include the increasing but unnoticed standardization of human actions, the silent loss of freedom and individuality, the increasing dependence on the smooth functioning of digital infrastructures, the loss of the future as an open space, or the loss of reflection and learning opportunities due to unlimited acceleration. The possibility of such gradual disruptions poses several challenges to responsible research and innovation (RRI), technology assessment (TA), and ethics. These include epistemological issues (how to detect gradual disruptions at an early stage), ethical issues (how to assess and evaluate concerns relating to the precautionary principle, for example), issues of whether countermeasures should be taken, and issues of communication between irrational exaggeration and irrational trivialization. The final part of the paper will address possible gradual disruptions that can be attributed to both technical parameters and human behavior, and draw conclusions for TA and RRI.

Keywords: Disruption, Digital Twin, Technological Dependence, Loss of the Future.

DOI: 10.17323/2587-8719-2025-4-68-83.

[*] Armin Grunwald, Dr. rer. nat., Dr. habil. in Philosophy (venia legendi); Professor at the Institute of Philosophy, KIT (Karlsruhe, Germany); Head of the Institute for Technology Assessment and Systems Analysis (Karlsruhe, Germany); Head of the Office of Technology Assessment at the German Bundestag (Karlsruhe, Germany), armin.grunwald@kit.edu, ORCID: 0000–0003–3683–275X.

1. INTRODUCTION AND OVERVIEW

Since the Second World War and especially in recent decades, technological progress has become a key factor in social development in many areas. Innovation and competitiveness require new technologies, such as in digitalization, medicine, or biotechnology, as well as for the transition to a more sustainable and climate-friendly society. However, this has led not only to the desired consequences but also to unintended, sometimes surprising, and often undesirable and problematic ones (Grunwald, 2019). These include major accidents in technical facilities (e.g., Bhopal and Chernobyl), the global environmental crisis (e.g., loss of biodiversity and climate change), stress for the labor market due to automation, risks for democracy due to problematic internet communication as well as the potential for dual use and misuse of technology at various levels. In current times, the divergence between intended consequences of technology and innovation and unintended ones, often manifesting themselves years or decades later, coincides with the emergence of a multi-polar world full of geopolitical tensions, including political competition in major areas of new technology like AI, robotics, quantum technologies, and biotech.

In this situation, *forward-looking* analysis and assessment of technology impacts are essential, in terms of both opportunities and possible unintended negative consequences. This diagnosis inspired the introduction of technology assessment (TA) in the US Congress in 1972 as scientific policy advice on the design and impact of technology (Bimber, 1996). This was followed by the spread and diversification of TA. Three main fields of practice can be distinguished today (Grunwald, 2019):

- ◇ TA as scientific policy advice, e.g., at the German Bundestag (see below), addresses publicly relevant, generally binding aspects of technology that must be decided by policy-makers, such as safety and environmental standards, the protection of citizens, the guarantee of human and civil rights, or the priority setting in research funding and technology policy.
- ◇ TA to support public debate and opinion-forming systematically engages citizens and stakeholders in debates on future technology, frequently involves the mass media, and sees itself as an element of deliberative democracy at grassroots level, beyond the institutions of representative democracy.
- ◇ TA in direct technology design accompanies the research and development of technology at universities and in industry. TA's knowledge

of consequences is incorporated directly into the development of technology, e.g., in order to design technology in line with values such as sustainable development or privacy.

Technological consequences are not simply the consequences of technology. They depend not only on technical parameters but also arise from the interaction of technical properties and human behavior, for example, through use and consumption. TA is therefore ultimately not about technology as such, but about researching and shaping socio-technical interactions. For this reason, TA is necessarily highly interdisciplinary and must involve engineering, social sciences, and ethics in particular. This applies equally to the field of *responsible research and innovation* (RRI; Von Schomberg & Hankins, eds., 2019).

Unintended consequences of the digital transformation have a different character than those of many other technologies. While TA has often had to deal with environmental, health, or accident risks in its history, for example, in the context of nuclear energy, these types of risks do not play a central role in digitalization. Instead, fears are repeatedly expressed here that can be understood as concerns about *gradual disruptions at a societal level*, i.e., about slow developments that can nevertheless grow into upheavals with considerable potential for damage (Section 2). Such possible upheavals characterize the debate on digitalization (Section 3).[1] They pose specific challenges for TA and RRI (Section 4).

2. ON THE CONCEPT OF GRADUAL DISRUPTION

Disruption has only become a widely used term in the last ten years or so. Although the word's origin refers to rather unpleasant-sounding meanings (lt. *disrumpere*, "to burst, break, tear apart"), it entered contemporary discourse with a positive intention. Disruptive innovations, based on technological leaps or entirely new business models, are valued in innovation policy (Vera & Ramge, 2021). In contrast to incremental innovations based on gradual product improvements, disruption is aimed at fundamental upheaval intended to overturn market conditions that have existed for years or even decades within a short space of time. New market opportunities are then open to the winners (often called "disruptors"). Also, entirely new markets can emerge, as in the digital transformation exemplified by platform economies such as Amazon or eBay or in digital photography.

[1] This publication continues and deepens earlier work by the author (cp. Grunwald, 2025).

In this way, the historically older theory of disruptive technology (Bower & Christensen, 1995) was quickly extended to the field of disruptive innovation (e.g., Danneels, 2004), in some cases with considerable expectations. However, the term is controversial (Gans, 2017): "'Disruption' is a business buzzword that has gotten out of control. Today everything and everyone seem to be characterized as disruptive — or, if they aren't disruptive yet, it's only a matter of time before they become so." In this criticism, the concept of disruption is reduced to a synonym for success.

For some years now, *crisis phenomena* have also been referred to as disruption. The coronavirus pandemic and recent geopolitical tensions are considered disruptive events. Both have ended a long period of broad stability, at least in the Global North, and, according to widespread diagnosis, indicate the transition to a time of permanent crisis. The term disruption is used here to describe the breakdown of stable social conditions. In communication, catastrophic narratives often come into play, such as the fear of nuclear war, climate change as the end of the Earth's habitability, the end of democracy, or the collapse of the labor market due to massive automation. Expectations of stability, assumptions of continuity, and planning certainty are breaking down and making future prospects appear uncertain. Semantically, this points to the time structure of abruptly occurring events. Seen in this light, talk of *gradual* disruption seems conceptually absurd or paradoxical.

A closer look allows us to differentiate. Semantically, the term disruption shows two elements of meaning: on the one hand, the *breakdown* of previously stable relationships and, on the other, the *speed of* this breakdown. While the first element of meaning is etymologically inherent in the term, the second can be handled more flexibly. Time scales of breakdown are elastic. For example, the invention of printing in the late European Middle Ages is often portrayed as disruptive — historically, this disruption extended over many decades of diffusion into the societies of the time. So, *on the one hand*, breakdown and discontinuation can certainly take place over an extended period of time. They are then only referred to as disruption in retrospect, in a kind of fast motion so to speak, whereas they appeared to the contemporary witnesses as a gradual transformation. *On the other hand*, discontinuation and breakdown, even if they occur suddenly, can build up slowly over longer periods of time. Nevertheless, everything remains stable for a long time and disruption only occurs later. The latter are referred to as *incremental* or *gradual disruptions*. This is what this article is about.

Many examples of this type of disruption are known from the technical world, especially those involving material fatigue and wear. The daily stress on many technical objects, such as bridges or V-belts in cars, gradually leads to the degradation of materials and components. They still function reliably for a long time until the wear and tear reach a level where something fails from one moment to the next, so that, in the example chosen, the V-belt suddenly breaks or the bridge collapses without warning, as happened in Dresden in 2024. In the medical field, strokes and heart attacks fall into this category. Certain signs can be recognized in advance with some uncertainty, such as calcium deposits in arteries, but the event then happens suddenly and unexpectedly. People often ask afterwards whether one could have known about it beforehand. One example from the climate debate is the so-called *tipping points* (see Gladwell, 2000). Further warming could lead to self-reinforcing feedback effects that would have dramatic consequences in a short space of time, i.e., a disruptive effect.

The disruptive effect in processes of this kind is therefore inherent in incremental processes that are difficult to recognize. It can remain unrecognized for a long time and thus escape early intervention and prevention. At some point, however, it can lead to potentially far-reaching and sudden consequences. The tragedy of such gradual developments is that in the incremental course, serious disruptions may announce themselves gradually, but can then take place abruptly. With this semantic differentiation, the possibility *of gradual disruption* in digitalization and the mass introduction of AI is considered in the following.

3. DISRUPTIVE POTENTIAL OF TRANSFORMATION THROUGH AI

The term *gradual disruption* can be used analytically to address possible developments in the digital and AI transformation with damaging or even catastrophic potential. This is not about predictions but about *possible* developments and corresponding concerns. They can be found at different levels in the debates on AI and digitalization. The following series of examples does not follow an ordering principle and does not claim to be exhaustive but reflects facets of the current debate on AI and digitalization (e.g., Deutscher Ethikrat, 2023).

SLIPPING INTO DEPENDENCIES

Modern societies are already completely dependent on the smooth functioning of critical infrastructures such as the power supply (Petermann et al., 2011). This increasingly applies to digital infrastructures. If the internet were to fail, financial transactions would become impossible, the global economy would collapse, media communication would no longer be possible, medical diagnostics would be deprived of many established procedures, international logistics chains would come to a standstill, and much more. The increasing introduction of ADM (*automated decision-making*) systems is creating a dependency on AI-controlled systems, which, together with their *black box* character and lack of transparency, but also due to the psychological *automation bias* (Deutscher Ethikrat, 2023; Safdar et al., 2020), lead to increasing dependence on these systems in decision-relevant contexts such as the police and social services.

The gradual displacement of cash is a current example of the ambivalence of technical infrastructures. While cashless payment transactions were initially an *additional* option alongside cash as a convenience for businesses and private individuals, there is now a gradual transition to a world without cash (Ehrenberg-Silies et al., 2022). Cash is slowly being displaced, partly driven by consumer behavior and convenience, partly driven by political and economic incentives and regulation, with the argument that this could make the black market and illegal work impossible. Once cashless payment transactions have become fully established, as is already largely the case in some countries, the freedom of choice in payment options will have disappeared and, if the internet were to go down, no more shopping or payment transactions would be possible. Due to its gradually increasing dominance, cashless payment becomes a compulsion, accompanied by full dependence on technical systems.

Dependencies are not disruptions in themselves, but they carry their seeds. Dependencies that have become total are *latent disruptions.* As disruptions in waiting, they build up gradually through growing dependencies, but in an emergency, e.g., if digital technologies were no longer to function smoothly, they can have abrupt and possibly catastrophic consequences. However, relying on their unlimited smooth functioning and making the functionality and stability of modern societies dependent on it is ethically problematic. Unexpected hacker events, a collapse of state order, or severe economic turmoil could also affect infrastructures such as the internet and, in the worst case, render them dysfunctional. Complete dependence on digital infrastructures and platforms is likely to have been reached long ago — which means that modern societies are already operating in the mode of this latent disruption.

LOSS OF THE FUTURE AS AN OPEN SPACE OF POSSIBILITIES

Digital technologies are often regarded as synonymous with the future, much like nuclear energy optimistically was in the so-called atomic age of the 1950s and 1960s. However, digital technologies generally operate based on past data. For example, digital twins (see above) only ever mirror a world of yesterday, e.g., in that customer profiles can only be created based on past purchase and consumption processes. Digital twins basically only depict the *past* of their analog originals. Big data technologies can only evaluate past data and recognize past patterns. AI systems can only be trained on data from the past, as data from the future is not available. Even if AI and *big data* are used to create quantitative forecasts, these are based on pattern recognition using past data. Due to the indispensable reference to data, digital technology is inescapably fixated on past conditions. When data sets, digital twins, and correlations and patterns uncovered by AI are used to make predictions about the future, past conditions are carried over into the future, imposed on it, so to speak. The future as an at least partially open space of alternative paths and possibilities is replaced by a data-based extension of the past.

Digitalization or some of its fields could become conservative in this way, aligning concepts for the future with old data rather than developing new ideas. Multiple anthropological determinations understand humans as beings with a future and the ability to envision and reflect on possible futures (e.g., Kamlah, 1973) — futures that go beyond extending the past into to the future instead include creative ideas in an open space of possibilities, which may even have a counterfactual and utopian character. A gradual disruption could occur here if the fundamental openness of the future fades into the background or disappears completely in favor of a data-driven orientation that remains bound to the past.

GRADUAL DISAPPEARANCE OF FREEDOMS

Human freedom again and again leads to unwanted effects. The example of road traffic, with over one million deaths worldwide every year, the majority of which are due to human error, is one example; crime and terror are others. Promises of security through prevention of accidents or defense against terrorism repeatedly provide arguments for interfering with civil liberties through surveillance and control. Regulation, the legal system, and security agencies should ensure that people do not exercise their freedoms at the expense of others. Technical surveillance and control systems are used to technically enhance security or to enforce it completely. Digitalization provides powerful tools for this (Spiekermann & Christl, 2016). Comprehensive surveillance by cameras, automated facial recognition, location tracking and creation of movement profiles, pattern recognition in offender profiles, technical requirements in operation, and even the removal of the "human factor" from technical processes, such as in autonomous driving, offer far-reaching opportunities to technically prevent the abuse of human freedoms — but also to abolish freedoms. Quite a few countries, especially in Southeast Asia, have achieved a high degree of technically implemented security and control in this way.

However, if security is enforced through technical means, there is no longer any freedom in the field concerned (Deutscher Ethikrat, 2023: 357). The tension between the high value placed on freedom, rooted in the European Enlightenment, on the one hand, and technical control and surveillance, e.g., to ward off terrorism, on the other, is a recurring theme in social debates on digital transformation.

The gradual disruption in this field would be an unnoticed slide into a world where the security interests of individuals and the state become the dominant value and are no longer weighed against other values, such as civil liberties. This would lead to ever-increasing control of human actions enforced by digital technology. Such a development would spell the end of individual freedom and erode the democracy based on it. The result would be a society controlled by digital means that is secure but completely unfree. Science fiction has repeatedly addressed such dystopian developments.

DIFFUSION OF RESPONSIBILITY INTO NOWHERE

Responsibilities are being redistributed at the constantly emerging new interfaces between humans and digital systems. Automated or autonomous decision-making systems (ADM systems), industrial production in cooperation between humans and robots in Industry 4.0, and autonomous driving are examples. However, the fact that machine systems are responsible for certain decisions does not mean that these systems also bear responsibility. This is because even AI-supported systems do not have intentions but merely use algorithms to perform complex statistical operations based on data. If they do not follow their own agenda and do not consciously want to achieve a specific purpose, they cannot bear responsibility (ibid.). The attribution of responsibility and accountability remains, at least for the time being and in the foreseeable future, the preserve of humans who act consciously and with intentions.

However, the attribution of responsibility to specific actors is becoming increasingly complex in a world with more and more AI systems. Although decisions and therefore responsibility in principle remain with humans, this is increasingly happening invisibly. While in a traditional car, the person driving is obviously responsible, this is much harder to recognize in autonomous cars. AI systems and their manufacturers interpose themselves between the intentionally acting humans and real effects, e.g., in the event of a traffic accident caused by an autonomous car. Responsibility shifts from individual drivers or, in the case of military drones, from soldiers to people and institutions in the background — to companies, programmers, managers, secret services, generals, or regulatory authorities.

Ethics and law have experience in assigning responsibility in complex contexts involving a division of labor, e.g., in large companies. The task of defining responsibility in constellations based on the division of labor between humans and AI systems also appears to be feasible in principle. However, the complexity of responsibilities distributed between humans and AI systems increases both the risk of a gradual "diffusion of responsibility" into nowhere and the risk of deliberate concealment of responsibility. In light of the philosophical ideals of linking freedom with responsibility, it is completely open whether and what kinds of possible gradual disruptions in the social order may result.

THE GRADUAL UNLEARNING OF ESSENTIAL SKILLS (DESKILLING)

In the course of historical and technological change, there are always skills that become dispensable and are forgotten. Examples of past professions that are no longer needed today can be found in museums. Knowledge is also lost, prompting historians and archaeologists to ask questions such as how the pyramids in Egypt or the Gothic cathedrals could have been built with the technical means available at the time. Such processes of forgetting take place slowly; new skills are developed in place of those that have been forgotten.

Digitalization, automation, and AI lead to similar, but massively amplified, more far-reaching, and accelerated effects. They make life pleasant and convenient in many ways, relieve people of the need to orient themselves in space through GPS, render learning superfluous in many fields, as knowledge is available digitally, and replace lengthy deliberations and the need to form an independent judgment through data-based calculations in ADM systems. Many fear that this could lead to the atrophy of abilities that are *constitutive* of being human (*deskilling*; cf. Deutscher Ethikrat, 2023: 353). As a result of the widespread use of AI applications, people could be increasingly tempted to delegate more and more tasks to AI technology because it is seen as supposedly superior or because it is convenient and saves time and effort.

An important role is played here by a psychological effect that is specific to AI systems: the *automation bias* (cf. Goddard et al., 2014). Many people trust algorithmically generated results based on large amounts of data and calculated using AI-supported decision-making processes more than those of human experts, no matter how much professional and life experience they have. The reason for this presumably lies in exaggerated attributions of objectivity and accuracy toward mathematical and data-based processes, on the one hand, and a suspicion of inaccuracy and subjectivity toward humans, on the other. Even if AI systems are strictly limited to *decision support* and human decision-makers have to make the decision, AI systems could gradually take on the role of the "actual" decision-makers and thus substantially erode human judgment.

In this way, key human skills and cultural techniques could be pushed into the background and eventually atrophy. Examples include the ability to understand complex texts when relying solely on short summaries generated by ChatGPT, or the ability to form one's own opinion on a complex issue when permanently and uncritically relying on the decision support of an AI application. The gradual disruption here would be the combination of the loss of human abilities such as judgment and critical thinking with an increasing dependence on AI systems.

END OF OPPORTUNITIES FOR REFLECTION AND LEARNING

Acceleration is part of the capitalist economic system. It unleashes creativity and innovation, primarily through competition. Acceleration is a phenomenon often discussed in the context of digitalization. The increase in computing speed, the possibility of calculating millions of options in the shortest time, the linking of creative resources via the internet, and the acceleration of data transfer and communication, much of it mediated and further accelerated by means of digital twins, all shorten innovation cycles. Accordingly, the above-mentioned "disruptive innovation" as extreme acceleration is the opposite of incremental innovation processes.

However, there is also destructive competition. The acceleration spiral is in danger of overexploiting the human and natural resources that feed it. One concern regarding AI-driven digitalization relates to negative and potentially ruinous consequences of ever-increasing acceleration, in particular to the question of whether and when further acceleration could fundamentally undermine important conditions of reflection. This would be contrary to the principles of enlightenment, the principles of technology assessment (Grunwald, ed., 2024), and the requirements of sustainable development.

Reflection requires careful analysis and deliberation, weighing alternatives, finding the right balance and ethically legitimate criteria for decision-making, as well as prudent implementation of the results, e.g., in legal regulation. All of this takes time in two ways: first, for the deliberation and consideration processes themselves and, second, for transferring the results into practical action and decision-making. The gradual disruption in this respect could be that societal capacities and structures for reflection would slowly erode under the pressure of capitalist competition. In the libertarian narrative of an innovation-oriented fatalism under the primacy of competitive thinking, reflection can no longer be afforded, since otherwise the competitor would be faster and gain market advantages.

4. REQUIREMENTS FOR TECHNOLOGY ASSESSMENT

Some of the developments described have already taken place (complete dependence on digital systems), some are observable (unlearning of skills), and some are only feared (diffusion of responsibility into nowhere). None of them are consequences of technology alone. Rather, gradual disruptions in connection with AI and digitalization arise from a combination of technical possibilities, business models, human behavior, and regulations. For example, AI does not actively take over the thinking for people, but people give up thinking for themselves and let “AI do the thinking.” Another example: dependence on AI systems arises from the fact that almost all routines in business and politics, but also in leisure and everyday life, now run via digital channels, and many are supported by AI. This is not a predetermined consequence of the existence of AI, but individuals, communities, or entire societies allow themselves to become so accustomed to these technologies in their behavior and habits that they are gradually becoming dependent. Digital technology and AI, together with applications and business models, provide the medium for gradual disruption, but are not solely responsible for it.

Digital and AI systems offer so many advantages that there is a pull toward their use and adaptation. In the process, digital systems are often overly trusted and people risk losing their own expertise (automation bias, see above). Undoubtedly, digital technology often makes life pleasant and convenient. As soon as routine activities at work or during leisure time have been adapted to digital systems, whether with or without AI, they are so much a part of life that it is often hard to imagine life without them, or at least it seems tedious and exhausting, hence unattractive. Such effects can also result from time pressure and efficiency requirements at work. If, for example, the individual examination of the records for processing an application for “Bürgergeld” (citizens’ income) requires working through extensive documents and takes time, while algorithms can do this quickly, provided the data is digitized, the willingness to let the algorithms do the work increases.

So it is not technology as such that leads to gradual disruption, but its combination with human behavior. When considering the consequences, the focus must therefore not be narrowed down to digitalization and AI as technology, but must instead take into account the interactions with human behavior. For TA, this is a fundamentally familiar but comparatively difficult constellation. The often only vaguely tangible human factor in terms of convenience, adaptation, and overestimation of digital systems — perhaps most strongly the “sweet temptation” of convenience — adds to the usual difficulties in recognizing gradual processes and assessing their relevance for action.

Gradual, creeping developments are often difficult to recognize at first. This is particularly true in their early phases, when insufficient data is available and only weak signals can be observed. The weak evidentiary basis, the lack of sensitivity to the only slowly developing potential for disruption, and the uncertainty as to whether a disruptive development will occur at all often reduce the willingness to deal with these developments proactively and, for example, to conduct empirical research to examine the evidence. Only when the signs of a disruptive development become more apparent does this willingness increase. Climate change as a structurally analogous, gradual disruption has provided illustrative material on scientific uncertainty and the growth of evidence since the 1970s. In digitalization, concerns about democracy were also initially rather speculative (Grunwald et al., 2006), whereas they have long since been empirically proven (Hofstetter, 2016). Also, the loss of competence due to the transfer of tasks to digital systems is no longer just a fear but has been substantiated by many examples in connection with the “ironies of automation” (Bainbridge, 1983).

In TA, the epistemological complexity is well known from conflicts over precaution, particularly in the health and environmental fields (Harremoes et al., eds., 2002). In typical precautionary situations, there is little knowledge about potential future damage and its probability of occurrence (Jonas, 1979; Von Schomberg, 2005). This epistemological challenge has direct consequences for the assessment and classification of developments that are only gradually becoming visible. The conclusion that TA and ethics should hold back until better knowledge is available (Nordmann, 2007) is out of the question in view of the high relevance of potential disruptions, in particular because of possible *points of no return*. However, prioritizations and urgency assessments require a certain level of evidence of knowledge about potential disruptions (Grunwald, 2010). A mere suspicion is not sufficient for a high prioritization, even if it would lead to a disastrous development if the suspicion were confirmed. There is the difficult task of assessing the situation and classifying it in comparison with other developments. The question arises as to when the evidence of a suspicion is sufficiently strong to at least legitimize the allocation of resources for more research in this area or even for intervening measures for preventive counteraction (Von Schomberg, 2005).

Due to the poor recognizability of gradual developments and the difficulties in assessing them, public communication about them is susceptible to ideology and speculation. On the one hand, there is a tendency to trivialize and downplay the issue, arguing that one should wait until better data and corroborated evidence is available instead of rashly wasting resources or unnecessarily restricting freedoms. On the other hand, weak signals are extrapolated into the future and dramatized to the point of stoking fears of rapid disruption. This results in mutual accusations of exaggeration, ideology, speculation, trivialization and whitewashing, as well as recklessness, irresponsibility, or permanent doubting. During the corona pandemic, these communication problems could be observed in many ways. Time and again, there seemed to be no path of mediating reason between dramatizing exaggeration on the one hand and downplaying the risks on the other.

Given the wide differences in the perception of opportunities and risks, there is certainly no one-size-fits-all solution to these communicative challenges. However, past debates on technology (Grunwald, 2011) show that neither trivialization nor dramatization are appropriate communication patterns. What is constructive is transparency and openness, including, and perhaps especially, with regard to the uncertainties of knowledge and the possible extent of damage.

REFERENCES

Bainbridge, L. 1983. "Ironies of Automatization." *Automatica* 19 (6): 775–779.

Bimber, B. A. 1996. *The Politics of Expertise in Congress: The Rise and Fall of the Office of Technology Assessment*. New York: State University of New York Press.

Bower, J. L., and C. M. Christensen. 1995. “Disruptive Technologies. Catching the Wave.” *Harvard Business Review* 69:19–45.

Danneels, E. 2004. “Disruptive Technology Reconsidered. A Critique and Research Agenda.” *Journal of Product Innovation Management* 21 (4): 246–258.

Deutscher Ethikrat. 2023. *Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz* [in German]. Berlin: Deutscher Ethikrat.

Ehrenberg-Silies, S., R. Peters, C. Wehrmann, and S. Christmann-Budian. 2022. *Welt ohne Bargeld — Veränderungen der klassischen Banken- und Bezahlsysteme* [in German]. Berlin: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag.

Gans, J. 2017. *The Disruption Dilemma.* Cambridge (MA): The MIT Press.

Gladwell, M. 2000. *The Tipping Point. How Little Things Can Make A Big Difference.* New York et al.: Little, Brown and Company.

Goddard, K., A. Roudsari, and J. Wyatt. 2014. “Automation Bias: Empirical Results Assessing Influencing Factors.” *International Journal of Medical Informatics* 83 (5): 368–375.

Grunwald, A. 2010. “From Speculative Nanoethics to Explorative Philosophy of Nanotechnology.” *NanoEthics* 4 (2): 91–101.

———. 2011. “Ten Years of Research on Nanotechnology and Society — Outcomes and Achievements.” In *Quantum Engagements : Social Reflections of Nanoscience and Emerging Technologies*, ed. by T. B. Zülsdorf, C. Coenen, A. Ferrari, et al., 41–58. Heidelberg: AKA GmbH.

———. 2019. *Technology Assessment in Practice and Theory.* Abingdon: Routledge.

———, ed. 2024. *Handbook of Technology Assessment.* London: Edward Elgar.

———. 2025. “Understanding the Digital Transformation. Philosophical Perspectives on Potentially Gradual Disruptions.” *Philosophy & Digitality* 1:3–13.

Grunwald, A., G. Banse, C. Coenen, and L. Hennen. 2006. *Netzöffentlichkeit und digitale Demokratie. Tendenzen politischer Kommunikation im Internet* [in German]. Berlin: edition sigma.

Harremoes, P., D. Gee, M. MacGarvin, et al., eds. 2002. *The Precautionary Principle in the 20th Century. Late Lessons from Early Warnings.* London: Sage.

Hofstetter, Y. 2016. *Das Ende der Demokratie. Wie die künstliche Intelligenz die Politik übernimmt und uns entmündigt* [in German]. Bielefeld: Bertelsmann.

Jonas, H. 1979. *Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation* [in German]. Frankfurt am Main: Suhrkamp.

Kamlah, W. 1973. *Philosophische Anthropologie. Sprachkritische Grundlegung und Ethik* [in German]. Mannheim: Bibliographisches Institut.

Nordmann, A. 2007. “If and Then. A Critique of Speculative NanoEthics.” *NanoEthics* 1 (1): 31–46.

Petermann, T., H. Bradke, A. Lüllmann, et al. 2011. *What Happens During a Blackout. Consequences of a Prolonged and Wide-Ranging Power Outage.* Norderstedt: BoD — Books on Demand.

Safdar, N. M., J. D. Banja, and C. C. Meltzer. 2020. “Ethical Considerations in Artificial Intelligence.” *European Journal of Radiology* 122.

Spiekermann, S., and W. Christl. 2016. *Networks of Control — A Report on Corporate Surveillance, Digital Tracking.* Wien: Facultas.

Vera, R. L. de la, and T. Ramge. 2021. *Sprunginnovation. Wie wir mit Wissenschaft und Technik die Welt wieder in Balance bekommen* [in German]. Berlin: Econ.

Von Schomberg, R. 2005. “The Precautionary Principle and Its Normative Challenges.” In *The Precautionary Principle and Public Policy Decision Making*, ed. by E. Fisher, J. Jones, and R. Von Schomberg, 161–165. Cheltenham: Edward Elgar.

Von Schomberg, R., and J. Hankins, eds. 2019. *International Handbook on Responsible Innovation. A Global Resource.* Cheltenham: Edward Elgar.

Grunwald A. [Грунвальд А.] Artificial Intelligence: Responsible Innovation in the Face of Potential Gradual Disruptions [Искусственный интеллект: ответственные инновации перед лицом потенциальных постепенных дисрупций] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 68–83.

Армин Грунвальд
д. филос. н., профессор
Институт философии Технологического института Карлсруэ (Карлсруэ)
директор Института исследования последствий
и анализа технологий и систем (Карлсруэ)
руководитель Бюро по оценке технологий при Немецком бундестаге (Карлсруэ);
ORCID: 0000–0003–3683–275X

Искусственный интеллект: ответственные инновации перед лицом потенциальных постепенных дисрупций

Получено: 13.08.2025. **Рецензировано:** 19.10.2025. **Принято:** 18.10.2025.

Аннотация: Данная статья рассматривает возможность постепенных дисрупций на уровне общества в целом в ходе стремительной цифровизации и распространения искусственного интеллекта (ИИ). Термин «дисрупция» означает внезапный распад привычных, ранее стабильных структур. Ожидания стабильности, предположения о преемственности и надежность планирования рушатся, окутывая будущие перспективы неопределенностью. Латинские корни этого термина означают «разрыв», «разлом» и «разрубание», семантически отсылая к временной структуре более или менее внезапных, резких событий. В этом свете упоминание о постепенной дисрупции в названии данной статьи кажется концептуально противоречивым или парадоксальным. Однако в мире технологий существует множество примеров дисрупций, которые были предварены заметными, но часто остававшимися без внимания признаками, в частности усталостью материалов и износом. Ежедневные нагрузки на многие технические объекты, такие как клиновые ремни в старых автомобилях или мостовые конструкции, постепенно приводят к их износу и деградации. В этом смысле понятие постепенной дисрупции отсылает к потрясениям со значительным или даже драматическим потенциалом ущерба, которые происходят не неожиданно и внезапно, как глобальная пандемия или землетрясение, а нарастают постепенно, пока, наконец, не приводят к разрушению ранее стабильных структур. В статье утверждается, что подобный тип потенциальной и постепенной дисрупции может возникнуть в сферах цифровизации и ИИ. Примерами служат растущая, но остающаяся незамеченной стандартизация человеческих действий, тихая утрата свободы и индивидуальности, растущая зависимость от бесперебойного функционирования цифровой инфраструктуры, потеря будущего как открытого пространства или утрата возможностей для рефлексии и обучения из-за безграничного ускорения. Возможность таких постепенных дисрупций ставит ряд вызовов перед ответственными исследованиями и инновациями (RRI), оценкой технологий (TA) и этикой. К ним относятся эпистемологические проблемы (как обнаружить постепенные дисрупции на ранней стадии), этические вопросы (например, как оценивать опасения, связанные с принципом предосторожности), вопросы о необходимости принятия контрмер, а также проблемы коммуникации между иррациональным преувеличением и иррациональной тривиализацией. В заключительной части статьи будут рассмотрены возможные постепенные дисрупции, которые можно объяснить как техническими параметрами, так и человеческим поведением, и сделаны выводы для TA и RRI.

Ключевые слова: дисрупция, цифровой двойник, зависимость от техники, потеря будущего.

DOI: 10.17323/2587-8719-2025-4-68-83.