



Beyond optimization: power, rationality and ethical risk assessment of AI

Florian Richter¹

Received: 24 August 2026 / Accepted: 26 September 2026
© The Author(s) 2026

Abstract

Emerging technologies pose various ethical challenges, including the need to assess their risks adequately. Still, the roles of different forms of rationality and power in assessing risks have not been fully scrutinized. To address this, we must differentiate forms of rationality to conceptualize power: instrumental, technical, and probabilistic. The latter has gained more power through the rise of artificial intelligence (AI), focusing mainly on optimization and accuracy. These forms of power and rationality lead to different conceptualizations of risk: real, hypothetical, and predictive risk. The ethical issues surrounding AI have not yet been fully scrutinized, since they are mainly approached through a lens of probabilistic rationality. I argue that optimizing algorithms to increase accuracy or engagement distracts from assessing whether AI-based systems are adequate for certain tasks. Furthermore, I discuss ways to govern AI-based systems through participatory processes and standardization that enable control and autonomous judgment in the deployment and use of such systems.

Keywords Autonomy · Artificial intelligence · Technology ethics · Control

1 Introduction

Emerging technologies pose manifold ethical challenges. The concept of power, however, is not explicitly discussed, and without a comprehensive understanding of power, ethical considerations in the field of artificial intelligence (AI) are futile. Different conceptualizations of power are rooted in different forms of rationality: instrumental, technical, and probabilistic. The one who determines what is rational also determines what is considered risky. I will argue that instrumental and technical rationality can help to set standards that are ethically acceptable and to a certain degree transparent. However, probabilistic rationality, particularly in predictive models, obscures the adequate establishment of

standards by focusing mainly on the accuracy¹ of performing specific tasks. It has been proposed that technology [18], pp. 71–73) and AI-based systems (van [61], Umbrello & van de Poel, [60]) should be aligned with values. The problem arises when the focus is mainly on improving accuracy (as in probabilistic rationality) without undertaking the irksome conceptual work of operationalizing values and treating the value trade-offs that ethical issues entail. The paper proceeds as follows: First, I describe the challenges of tech ethics to frame the problem. Second, I will develop the forms of rationality and the corresponding conceptualizations of power. Third, I will discuss the ethical implications of these conceptualizations for risk assessment.

✉ Florian Richter
florian.richter@ku.de

¹ School of Transformation and Sustainability, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany

¹ Cooper & Abrams state that the trade-off between accuracy and fairness is mainly the underlying practical assumption of much of the literature doing empirical work, where the trade-off is taken as “inherent,” and an “increase in fairness necessarily comes with a decrease in accuracy; increasing accuracy necessarily decreases fairness.” However, they also argue that the “normative assumptions” must be made explicit. For instance, if a model relies on “past unfairness” encoded in the data, increasing accuracy will lead to advantages for privileged groups. [10]. They point to the problem; however, they cannot resolve the tension since they do not locate the issue within a certain form of rationality and compare it to other forms for approaching such issues.

Table 1 The different forms of rationality, power and risk

Forms of rationality	Forms of power	Forms of risk
Instrumental rationality	Conditioning power	Real risk
Technical rationality	Regulating power	Hypothetical risk
Probabilistic rationality	Predictive power	Predictive risk

1.1 Challenges of technology ethics

The paradigm of classic technology is to control and regulate processes to ensure “repeatability” [20, 53]. In so-called “transclassic” technology, we face an “intellectualization of the environment,” for example, in ubiquitous computing, i.e., access to the world is mediated via technology and intelligent systems (for example, via sensors, cameras, etc.) [29], p. 43). Technology ethics faces several other challenges, e.g.,

- The increase in complexity of intelligent systems that makes it more difficult to operate and control them,
- A higher depth of engagement through, for instance, chips that are implemented into the brain and cannot be easily disentangled from the human body,
- And the interaction with autonomous systems or virtual agents.

Furthermore, generative AI shifts the paradigm of human-system interaction, moving it toward human-system co-action. Ethical assessment of human-system interaction, already disrupted by AI technologies, may face new challenges in developing assessment tools for human-system co-action [51]. This is even more pressing since the conceptual framework for handling risk and uncertainty still needs to be adequately developed for AI technologies. Philosophical approaches have focused on how our access to the world and ourselves is mediated via technology [54, 62] and how AI technologies may manipulate users [36]. Furthermore, the concept of power has been examined in technology ethics [32] and AI ethics [45]. However, no comparison or systematic analysis of the underlying conceptualizations of power and rationality has yet been made.

This paper makes three contributions. First, power is rooted in different forms of rationality. Power is an enigmatic concept in philosophical and political discourse. However, technological developments fall under different conceptualizations of power. Hence, the concept must be scrutinized from a technologically informed perspective; otherwise, pressing ethical questions, such as those regarding AI, cannot be understood. Second, I discuss counter-strategies against the exercise of power, as well as forms of legitimizing power in technology development. Third, this results in conceptualizations of risk that are aligned with

the forms of rationality to evaluate the implications of (AI) technologies.

Power can be exercised in different forms. Behind them stand different forms of rationality that shape this exercise (see Table 1). As argued here, power is rarely related to a specific form of rationality or to other forms, which makes it ultimately difficult to assess the risks that spring from different forms of rationality and power. Technical rationality focuses on securing the repeatability of actions. For that reason, certain standards or choice architectures must be in place that shape the environment to ensure reproducibility. For instance, the annoying noise that reminds car drivers to fasten their seatbelts shapes user behavior and reduces the risk of severe injuries. The (hypothetical) risk can be assessed in crash tests (experimental situations) to demonstrate the effectiveness of the safety measure. Probabilistic rationality stands behind predictive power. For instance, people can be placed in certain risk categories for the likelihood of repaying their loans. Whilst test scenarios in experimental settings can, to a certain extent, establish causal relationships, probabilistic rationality shows only correlations, where performance accuracy can be improved. A trade-off between accuracy and fairness seems, from this perspective, inevitable. All forms are, however, different respects of approaching ethical issues.

1.2 Philosophical approaches to power

In the early modern period, questions of power were primarily discussed in relation to the state’s legitimacy, as in social contract theories. However, modern thinkers have also examined power in relation to socio-technical systems and their effects on humans, for instance, Max Weber, Michel Foucault, and Byung-Chul Han.

(1) Instrumental or conditioning conception of power:

Max Weber included the concept in his sociological studies, where “power means every chance to assert one’s own will within a social relationship, even against resistance, no matter what this chance is based on” [63], p. 28). Although he states that the concept is “sociologically amorphous,” he concretizes it in the concept “dominion,” which is the opportunity to find obedience for an order [63], p. 28). Power is the chance or opportunity of diminishing resistance either by possible gratifications or sanctions. The one who does not fear the latter or is not interested in the gratification is not “controllable.” It is a modal concept and does not directly materialize; otherwise, it would break down. Someone is controlled or under the power of someone by possible sanctions, not by direct sanctions, since this would be (at least in some cases) violence to move someone, not anymore power. It is a “hypothetical force” based on the effectiveness of the condition (gratification and sanction).

[32], p. 17) Hence, it is an instrumental conception of power that also sees discipline as rehearsed behavioral schemata [63], p. 28/29).

(2) *Regulating conception of power:*

For Foucault, the effects of power should not be described in “negative terms” [15]. Those effects stem from a “heterogeneous totality, consisting of discourses, institutions, architectural facilities, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral, and philanthropic tenets” [17], p. 392). For instance, individuals are assigned to places, and schedules are set to organize their days. This happens in factories, army barracks, schools, universities, hospitals, etc., to organize the lives and bodies of individuals. Individuals are just instruments of power techniques; their bodies are made “docile” and disciplined. The performance of the individual is measured against a standard, for instance, in school exams. This “power of normalization” enables both homogeneity and individual differentiation [15]. The students should be able to read and write according to a set standard; some might excel more and others less.

Although it does not result in the subjection of the individual or in instrumentalization, as in conditioning power, but rather in alignment with a norm [55], p. 94), it is seen as a subtler version of the exercise of power that permeates life. Its normative force is not directly connected with (possible sanctions or gratifications) but with the adjustment to norms and ways of life. Freedom is the “condition” of power since power without the opportunity to resist would be “physical determination” [16], p. 790). Alignment to certain social norms, like going to church, can be resisted by bearing certain social costs. It is a normative, not a causal, force that prompts someone to go. However, it must not be understood as something purely negative since modern state power is also concerned with disciplining the individual for the sake of safety and health [55], p. 88). The individual should be enabled to participate in society as a mature citizen, exercise a profession, etc. All of this presupposes a kind of discipline. Foucault demonstrates how power is performed in disciplining citizens. Nevertheless, describing power solely in performative terms is insufficient, since the more essential part is the public discourse about ethical issues, the reasons for using certain disciplining forces, and how we set criteria for achieving public safety and health. For instance, compulsory vaccination can be reinforced by allowing only vaccinated children to attend public educational institutions. Such performance of state power might be understood as an invasion of the citizens’ freedom to decide over the bodily integrity of their children. However, the state also has an obligation to ensure the well-being of its citizens (based on technical rationality).

(3) *Seductive conception of power:*

Another conception stems from Han, who speaks of “smart power” [21]. However, the term “smart” has been used ubiquitously in recent times and has become somewhat of a buzzword. Since he characterizes this conception as not “repressive” but “seductive” [21], it could also be called “seductive power.” Social media platforms exemplify such a form of power. They seduce the user to spend more time on the platform, so they are not violent, “prohibitive,” or “censoring.” According to Han, such power is “*permissive*” since it is rooted in “friendliness.” While politics is about “passive surveillance,” what he calls “psychopolitics” is about “active control” [21]. In Foucault’s conception of power, the state takes the active role of disciplining. However, according to Han, platforms, or social media, shape choice architecture in ways that influence or seduce users into staying engaged.

This rootedness in friendliness is evident in Facebook’s algorithmic adaptation, which prioritizes posts from family and friends, thereby “increasing user engagement” [1]. Hence, this form of power presents itself as being mainly seductive. However, it is unclear exactly how to address the risk that such a form of power poses, as, for instance, according to Acemoglu and Johnson, “misinformation and political polarization” were also amplified [1]. The risk assessment of, for instance, social media platforms must be grounded in a technical rationality based on scientific studies (see the end of Sect. 4.2).

(4) *Critique of the philosophical conceptions:*

According to Hubig, Foucault presents “heuristic metaphors” [32], p. 116) and is only “demonstrating the power process in its performance” without theoretical and conceptual foundations. Hubig elaborates this conceptual foundation for the regulating conception by explicating technical rationality and abductive inferences [28]. This critique can also be extended to the conceptions of Weber and Han, as they do not ground their conceptions explicitly in types of rationality, namely, instrumental and probabilistic. This theoretical grounding will be presented in Sect. 2 to show how different forms of risk emerge from different forms of rationality. Demonstrating the performative aspect does not lead to ethical assessment, since it is unclear whether certain power mechanisms are considered ethically acceptable. Compulsory vaccination programs can be considered ethically acceptable (regulating conception of power) to ensure public health, particularly when certain incentives or gratifications seem to fail (conditioning conception of power). Furthermore, this performative aspect of power must be complemented by questions of legitimacy (see Sect. 3.2) and governance (see Sect. 3.3); otherwise, power will be conceived as something overwhelming and coercive.

1.3 Risk, ignorance, and uncertainty

The concept of risk is accompanied by several other concepts that help determine its use or meaning, such as chance, uncertainty, security, damage, and benefit (for some essential distinctions, see 46, pp. 5–11)). Security, for example, is seen as the “inverse of risk” [43] or as the “absence of unreasonable risk” [33]. However, these interdependencies do not help determine risk more sharply. In decision theory, on the one hand, a “decision under risk” would be where the “decision maker knows the probability of the possible outcomes.” On the other hand, a “decision under ignorance” would be where “the probabilities are either unknown or nonexistent.” Uncertainty, however, is “either a synonym for ignorance, or as a broader term referring to both risk and ignorance.” [48] Nevertheless, it would make it impossible to handle cases where we cannot calculate risk and probabilities. According to Aristoteles, one decides about something that one can know in advance because of reasonable deliberation, but one decides about something unforeseeable according to one’s particular habitus (ἔξις) [4] (EN III 11, 1117a). Something unforeseeable cannot be known in advance or predicted by probabilistic means. It is a decision under uncertainty. One must acquire specific competencies to possess a particular constitution that enables one to deal with such uncertainties. Many of our decisions are based on this because we face uncertainties we cannot predict, such as choosing a partner or a job. We must trust ourselves, our decisions, and others. However, it must also be possible to inquire scientifically and conceptually about uncertainties. Hubig’s tripartite classification into real, hypothetical, and meta-risk offers advantages in sharply differentiating technology risks and in developing tools to handle uncertainty [27], pp. 80–92). It will be shown that AI models need to be assessed this way; otherwise, the risk arising from AI remains a vague term without any conceptual grounding.

Technologies and AI are not just systems that differentiate themselves from their environment; they also have some functional characteristics (following Niklas Luhmann broadly). They are modeled systems or architectures with different purposes and contexts. Cognitive architectures that model the functioning of the mind (either as symbolic or connectionist), computational architectures that implement cognitive architectures or AI models and calculate them (for example, in “simulation pipelines” for “virtual experiments” [9], p. 1/2) or AI pipelines), and choice architectures that model possible choices and, in the end, also allow a certain degree of automatization of decisions. Modeling involves selecting which parts, functions, principles, and performances are most relevant to understanding a system. Nevertheless, models have their (ontological and epistemological)

limitations that must be analyzed; otherwise, the risks involved will only be understood superficially.

1.4 The status of models

A crucial distinction between “model of...” and “model for...” must be considered in modeling [28]. For example, a city map is a model of a city. Exactness and verification depend on the crucial features needed to represent the city adequately. Nevertheless, an exact representation, such as a 1:1 model of the city, would be highly impractical or not even feasible. Adequacy is always dependent on the choices made. Ultimately, such a model can only be tested empirically in an experiment that demonstrates its validity by serving as a model for a specific goal. For instance, a city map is considered valid if it provides a clear orientation in the city. It should not be understood that one should opt for pure instrumentalism or pragmatism just because the model is helpful. Maps available online and in navigation tools are very helpful. However, the adequacy and truthfulness of the representation can be questioned. For instance, are the restaurant options an adequate representation of the city’s food options? Are some presented with more emphasis because the owners paid for that service, or did they/someone write favorable comments about the restaurants? These strategies are usually not apparent to the users of such services. Furthermore, goals are also embedded in a broader context of value decisions.

Generative AI models are also models of language, cognition, and the mind. Here, a specific cognitive architecture is taken as the basis (connectionism and deep neural networks) that should represent “adequately” the functioning of the mind to produce, for instance, text. Nevertheless, as a model for..., it must demonstrate its validity within specific contexts for particular goals. The often-problematized point that AI models might perpetuate biases because their database (input) is biased depends on the cognitive architecture: their “primitive” selection of inputs that are not evaluated, and their functioning through weighted probabilities. These inadequacies as a model of... are disguised by their instrumental success in facilitating translations or generating text. Translating a text for a foreign-language class to complete homework quickly is also embedded in value decisions that extend beyond pure usefulness or process optimization and may lead to problems later when such competencies are needed.

1.5 Choice architectures and personalization

Information is more and more technologically mediated. It is presented on social platforms and sorted by algorithms and decisions other people make by liking it, for instance.

Newspapers also gather information from sources, process it, and present it in the form of articles, so the information is also mediated technologically and in other forms. However, to a certain degree, the process is transparent, even though people have less and less trust in these institutionalized processes. In a way, social media platforms, newspapers, and other media represent the world in a way, so that people can make decisions like voting in politics, buying products, etc. It is a complex interplay among individual agents (citizens, consumers, etc.) and collective agents (governmental and non-governmental agencies, companies, etc.). Manifold choice architectures are designed, from complex ones, such as tax forms, to simpler ones, like the arrangement of products in supermarkets. Regarding such choice architectures, discussing concepts such as nudging, defaults, power, manipulation, etc., is essential. Nevertheless, it is crucial to analyze them concerning AI-based systems since the collection of large amounts of data combined with probabilistic or statistical models enables the prediction of behavior, whether it be of persons or things. This move is mostly done under the heading of personalizing services for individuals. It can be personalized learning paths, offers, etc. However, the strategies behind such personalization are often not transparent, neither for the user nor for the developer. So, while personalization is often sold as an advantage (like helping weaker students to obtain a fairer education), it is in fact the main problem since it lacks transparency, is an intrusion into privacy, and threatens the autonomy of the person due to offering choice architectures that might nudge the person to do something.

On the other hand, institutionalized frameworks offer transparent guardrails with opportunities to develop a critical stance toward them. Societal systems, such as educational systems, are modeled and institutionalized. They have norms, laws, rules, and requirements that one needs to follow to behave adequately in such systems. In the case of educational systems with subsystems like universities, there are module descriptions, study and examination standards/regulations, and so on to guide one's actions. In other systems, such as the health system, definitions are needed to have a common understanding of diseases or even what health is, as the World Health Organization (WHO) described in its constitution [64]/48). Processes, procedures, the selection of adequate means, and the options to set specific goals must be transparent and written down for the systems' participants. Furthermore, certain forms of protest, options to abstain from particular actions, the development of moral judgment, the statement of free opinion, etc., must be allowed, although sometimes only to a certain degree.

Here, the concept of power comes into play because it needs to be critically reflected on whether protest is possible without a violent counterreaction from those in power, but

also whether someone is nudged in the selection of means, even though there has not been a consensus on whether all share the goal. Thaler and Sunstein propose such a "paternalism of means" [58], p. 7 and 312). At first sight, technological systems seem to be adequate means to nudge people's behavior. In smart environments, nudging might seem viable for optimizing not only technology use but also human behavior. However, in their "paternalism of means," Sunstein and Thaler suppose a common-sense understanding of morality to which people should be nudged. Ultimately, it is not only a paternalism of means but also of goals. Moreover, by excluding users or citizens from deliberating on normative standards, it undermines democratic deliberation, citizens' and users' liberty, and autonomy at both the strategic and moral levels. Developing and deploying technological systems that make decisions at a strategic or moral level must be accompanied by an assessment of whether stakeholders share the goals. Such strategies need to be made explicit [50]. It is an exercise of instrumental rationality in which technology is a means to reach goals.

2 Forms of power and rationality

Although choice architectures are designed to guide, steer, and regulate human behavior, on a normative level, certain forms of rationality are assumed to orient their construction. Assuming a specific form of rationality allows for the deconstruction of the adequate resistances that steer behavior. For instance, instrumental rationality is guided by maximizing utility through optimizing the tools to achieve it. The more useful the tool is, the more likely it will be used.

Choice architectures are, in a way, the aforementioned "heuristic metaphors" that demonstrate the power process in their performance. However, power is rooted in different forms of rationality; otherwise, on the one hand, the risk of power could not be assessed, and on the other hand, its adequacy could not be evaluated. For instance, nudging someone to eat healthier is based on instrumental rationality, i.e., choosing the proper means (food). However, it might not be adequate to nudge a person with an eating disorder since it might worsen her sickness.

Control and regulation can be differentiated into two main forms of power: conditioning and regulating power, both used to discipline people (see Weber and Foucault above). They are rooted in instrumental and technical rationality. Yet, AI-based systems offer another form of power: "predictive power" [45]. This form is partly used in AI-based systems that employ statistical algorithms learning from data. It is rooted in probabilistic rationality. I will argue that such algorithms are used to predict behavior and personalize smart or intelligent environments that, on the one hand,

deconstruct resistance to the world or to others, and, on the other hand, deprive individuals of opportunities to assess the risks of their decisions. Dating applications ease contact by reducing the difficulty of initiating a conversation in a bar. A computer game, where purchase of additional items might be handled in the currency of the game, can still be based on real transactions when the credit card is connected to the player's account. The virtualized context obscures financial transactions and may facilitate the purchase of many virtual items. Also, personalization may reduce the chance to assess risk. In a smart environment, such as a smart supermarket, monitoring food prices can be difficult. For a person approaching items, prices might change based on her past purchases. The customer might no longer be able to assess why she pays the price she does. Normally, prices across different supermarkets can be compared, and prices might reflect market "logic," such as shortages of specific products leading to higher prices.

2.1 Conditioning power—instrumental rationality

Taking the *homo oeconomicus* as a "means for a purpose" [2, p. VIII) yields modeling decisions as instantiations of a certain rationality or rules that a rational agent *should* follow. An agent that does not act by this rational ideal is then considered "irrational" and "not compatible with the model of the *homo oeconomicus*" [2, p. VIII). Such models are normative and not empirically contentful. That is why the "goals and needs of the involved cannot be derived from the decisions" [2, p. VIII). The aims are presupposed and are "rational aims" as long as an agent considers them worthwhile; only "sets of aims can sometimes be irrational, e.g., if they are mutually inconsistent" [48]. Decision theory is mainly about "instrumental rationality" and thus choosing the proper, optimal, or adequate means to accomplish the goal [48]. The goals and objectives one aims to achieve are those considered useful, worthwhile, or gratifying. In the end, they are those that increase utility or maximize happiness. The above-mentioned "paternalism of means" by Thaler and Sunstein operates within such a framework [58], where goals are assumed to be good for the user, since she is nudged, for instance, to eat healthier.

Conditioning power is based on a behavioristic framework. Control is exercised by positive or negative reinforcement. It is a direct form of instrumentalization. Gratifications can also be postponed in order to obtain them later. Exercising control and power can be technologically mediated, as in social credit systems implemented and enforced by AI-based systems and other surveillance technologies. In that sense, it aligns with Weber's conceptual concretization, in which "dominion" is the opportunity to secure obedience to an order [63, p. 28). Power is the chance or opportunity

to diminish resistance by offering possible gratifications or threatening with sanctions [32, p. 17). Power struggles can be boiled down to conflicting interests [28, p. 172), where some goals might be taken as gratifying. The one who does not fear sanctions or is not interested in the offered gratification is not "controllable." It is a "hypothetical force" based on the effectiveness of the condition (gratification and sanction) [32, p. 17).

Based on the collected data, it is possible to estimate the risk that gratification will be refused or that the sanction will be ignored. This helps to optimize control. For calculating risks, see also Sect. 4.1.

2.2 Regulating power—technical rationality

Instrumental and technical rationality are commonly mistaken, since instrumental rationality involves using tools to achieve goals. However, this is only one aspect of technology. Using a tool is an open-loop control system in which some input variables influence the system to alter output variables; for instance, a light switch is used to turn on the light. Regulation, as defined in control theory, means that an actual value is continuously aligned with a target value through a (closed) feedback loop; for instance, a thermostat measures the room temperature and controls the heating to align it with the target value. In a (closed) feedback system, "perfect regulation enables successful control" [6, p. 290).

These are aspects of technology as a concept of reflection, as argued by Hubig [31] and Grunwald & Julliard [20], where it is about the "achievement of situational invariances in the suitability of means." [20, p. 143) This is probably why Ropohl also mentions "reproducibility" as the aspect of the reflective concept "technology" in Grunwald and Julliard [53, p. 50). Situational invariance is not only about the optimal use of means (control), but also about regulating and securing the use of resources [31]. According to Hubig, a "common interest" of human technology—"in addition to the realization of purposes in general"—is therefore the "safeguarding," i.e., the "realization of purposes qua repeatability, plannability, anticipability" [31].

Experiments testing hypotheses establish situationally invariant relations between phenomena by demonstrating a significant influence of the independent variable on the dependent variable. Experiments should be reproducible within laboratory settings. This is the basis for real-world applications of the established relations, helping make processes plannable and repeatable. Technical rationality encompasses inductive and abductive reasoning to uncover those relations. Furthermore, standards are established as normative frameworks to guide technological development and application; on the one hand, to ensure functionality and

interlocking of processes, and on the other hand, to ensure ethical alignment.

Regulating power seeks, for instance, to exercise control over groups through human population planning or control to ensure that behavior is predictable and controllable. Regulating power should have a disciplining effect, sanctioning behavior that deviates from established norms, where the individual's performance is measured against a standard, for instance, in school exams. This "power of normalization" enables both homogeneity and individual differentiation [15]. The disciplining does not result in the subjection of the individual or in instrumentalization, as in conditioning power, but rather in alignment with a norm [55], p. 94). For that purpose, at the systemic level, across the health, legal, and educational systems, etc., the functioning must be ensured by offering transparent standards/regulations. In the health system, the World Health Organization's "International Classification of Diseases" (ICD) serves as a standard for identifying diseases. However, it may be a problem for individuals when their symptoms or suffering do not fall into this classification. This can lead to incorrect treatment and non-coverage by health insurance for treatments. In technological development, norms via standards are established by organizations such as ISO and IEEE to ensure that devices and systems operate as expected and minimize risks. In legal systems, laws are legislated to ensure public safety. For the discussion of risks emerging from regulating power, see Sect. 4.2.

2.3 Seductive and predictive power—probabilistic rationality

AI algorithms, particularly those based on deep neural networks, have achieved great success in performing tasks accurately, ranging from image recognition for cancer detection to image, video, and text generation.² Nevertheless, many ethical issues arise in the use, development, and deployment of such systems. For instance, Merchant et al. were able to predict health conditions from updates of social media users [44]. This is a double-edged sword since it might, on the one hand, allow the prevention of, for instance, suicides by detecting suspicious posts or likes on social media. On the other hand, it might be used by insurance companies to deny health insurance to high-risk patients. Kosinski and Wang have shown that image recognition systems can be used to detect the sexual orientation of persons, and even more

accurately than humans [37]. The problematic issue is that the focus is mainly on accuracy as a paradigm of truth and not adequacy [49]. Here, a kind of "predictive power" [45] is used to make "accurate" predictions about the behavior or status of persons.

Mühlhoff discusses this form of power with regard to "predictive rationality" and its underlying Bayesian conception of probability (which he distinguishes from a frequentist conception of probability) [45], pp. 121–132 and 160). The distinction goes back to Breiman, who divided statistical modeling into two cultures when trying to understand the connection between input and response variables. In one paradigm, a data model (or hypothesis) is established. It is validated using logistic or linear regression to determine whether the relationship is statistically significant. In the other paradigm, the prediction of the response variable is performed using decision trees or neural networks and validated by its "predictive accuracy." The so-called "algorithmic modeling" is solely an "emulation of nature" or of the processes in question. [8] The emulation of tasks, for instance, image recognition, can be validated by its accuracy and then compared to the accuracy of humans performing the same task.

Yet, the use of predictive models comes with a "qualitative leap," since it is not only possible to make predictions about users who explicitly make their data available, for instance, on social media platforms, but also about users solely based on their behavioral data on the platform (likes, posts, etc.). A person who does not disclose her age might still receive targeted advertisements based on her age, which might be inferred from her behavioral data. [45], p. 94/95) Such predictive models are also used in so-called "predictive policing" or for predicting the risk of criminal recidivism. According to Mühlhoff, such cases present non-trivial issues in which we could weigh "predictive privacy" against a value such as "public safety." He believes such an intrusion into our privacy cannot be morally justified, as it dismisses the exercise of free will by the affected people. [45], p. 109/110).

Emulating processes to yield accurate predictions does not imply a good understanding of the underlying processes. This is, in a general sense, the critique of Breiman's "algorithmic modeling," which he favors, that "success at prediction being the criterion of success, as contrasted with issues of interpretation or understanding." For a good understanding, hypotheses must be tested. They spring either from empirical investigations or theories [11]. Another critique that can also be found in machine learning models is overfitting: "High performance on the test sample does not guarantee high performance on future samples, things do change" [25].

² Probabilistic rationality is taken here as a more general form for different statistical algorithms and cannot be equated to machine learning prediction or deep neural networks. (I am thankful to an anonymous reviewer for pointing this out). Forecasting a word while producing a text is different from forecasting an event. Nevertheless, the same rationality underlies them.

Furthermore, the performative aspect of power is here neglected. As described above, Han conceptualizes power as “seductive” [21]. However, framing it mainly within “predictive rationality” will lead to an inadequate understanding of (harmful) consequences, because the prediction of behavior might be used to improve targeted advertising. A scientific and risk assessment must be grounded in technical rationality, as it enables the development of evidence-based conclusions; for instance, regarding health problems related to social media use (see 4.2).

In summary, Mühlhoff grounds predictive power rightfully in a specific form of rationality. However, he misses the seductive aspect of this power concept, which fuels the input of such models. Additionally, Mühlhoff criticizes the cybernetic nature of “Human-Aided AI” [45], pp. 135–145). However, the technical rationality is also based on cybernetics since it tries to regulate control. This should not be seen as a bad thing. In technical rationality, it is possible to identify value conflicts and make the criteria for operationalizing such values explicit. And in the end, the regulation helps ensure their consistent deployment. Nevertheless, he is right about the opacity of predictive power in the implementation of values and the largely inadequate deployment of predictive models in social, educational, and economic systems. The assessment of their risk will be discussed in 4.3.

3 Counterstrategies, legitimacy, and governance

In recent years, the goal has been to align AI-based systems with values. Such a programmatic idea is discussed under the heading of Value Sensitive Design [18], pp. 71–73), which has also been adopted for AI-based systems and how they can embody values (van [61], Umbrello & van de Poel, 2021). Others have proposed designing AI for the social good. [14] However, the affected are users, and, for that purpose, they should also be included in the development process, so that the legitimacy of using AI-based systems is sought through democratizing AI and broadening participation in its use, development, and governance [42]. Another strategy would be for users to seek counterstrategies or subversive uses.

3.1 Counterstrategies—subversion

Users or customers could use AI to improve their negotiation strategies with companies [41]. Also, Foucault mentions the strategic aspect of resistance [16]. Nevertheless, there is another form of resistance to consider when dealing with power. Such a form is subversive, creative, and more a part of a way of life [55], p. 134). In

youth cultures, different forms of resistance have been used, for instance, to appropriate inner-city spaces for skateboarding [3]. Such forms of resistance can also be channeled by building skate parks to get young people out of the inner city. However, for the use of AI-based systems, new forms of creative, subversive resistance must first be developed.

3.2 Legitimacy—participation

Under the heading “democratizing AI,” approaches are developed that hope to mitigate algorithmic and socio-technical injustices. This could mean either democratizing the use through broader accessibility of AI, the development through participatory research, or the governance through democratic institutions [42]. It has also been examined which role AI itself may play in improving or inhibiting “the democratic nature of political systems” [47]. Others have argued that just “deepening participation” is insufficient and there must be an increase in “the democratic quality of the administrative and executive elements of collective decision making.” Mere broadening of participation might lead to a “decrease in (the quality of) deliberation.” [24] Moreover, the “majority rule” might entail excluding “the voices of the marginalized.” Therefore, Lin proposes “*deliberative mini-publics*” as another option. Such forums consist of “randomly selected populations” in order to “deliberate with other participants on issues of public concern, and derive results that can be used as guidance for public decisions [42].”

Participatory research is another option, as it allows testing hypotheses through behavioral experiments. For this purpose, the difference between mere acceptance and acceptability, which is based on reasons, must be clarified [29], p. 162, [19]. Who would object to a widely accepted claim that eating healthily is good? However, which normative or moral factors play a role in considering certain food choices acceptable or not? Another approach would be to reveal preferences through controlled behavioral experiments, for instance, to determine which fairness standard to use for generative AI [52]. Nevertheless, it has been pointed out that this could lead to “participation washing” [7], where participatory approaches are used to legitimize pre-determined goals.

Mainly, big tech companies develop AI-based systems. And even though they set ethical principles to guide development, those guidelines remain fairly abstract [42]. The regulatory aspect of operationalizing values or balancing value conflicts is crucial since it affects the core of aligning AI-based systems with values. The problem, however, is which kind of trade-offs are considered. Lin,

for example, asks: “How can a trade-off between values, such as accuracy and fairness, be made?” [42] For this trade-off, see also 4.3.

3.3 Governance via standards

In AI governance, the focus is mostly on state or supranational institutions, as for instance discussed by Lin [42]. The usual criticism of such regulations is that, on the one hand, they take time to develop and can quickly become outdated by new developments, and, on the other hand, they tend to overregulate, thereby slowing development in those areas. However, other forms of organizational standards, such as those from ISO or IEEE, provide normative guidance for the development of new technologies. Here, a participatory opportunity for different stakeholders is also possible, as the guidelines are usually developed by interdisciplinary teams of researchers and practitioners. They stem from expertocratic deliberations and do not provide directly democratic legitimacy. Despite that, they are important tools for technical and normative guidance.³

Both institutional measures have the advantage of citizen and other stakeholder oversight of technological developments, not just big tech companies, and thus contribute to a kind of legitimacy by incorporating reflected value decisions that are accessible to users and open to critique. Every elaborated social system has a normative framework in which norms and rules are explicitly written down, and forms of protest are permitted. For instance, universities have norms and regulations for measuring student performance, and students can, within certain institutional limits, protest their grades. Deviation from the norm determines the grade, and the norms (correct answers for closed questions) must be provided to students. In technological development, ISO standards provide criteria for values such as safety when developing artifacts or systems to avoid unreasonable risk. Since AI-based systems can emulate human tasks well, such as driving a car or detecting cancer in images, their accuracy can be compared to that of a human driver or physician. More problematic are cases where the adequacy of such systems is in question, as with COMPAS, where risk scores are presented to judges to indicate whether a convicted person might reoffend. Here, the system’s accuracy against a human judge cannot be directly compared, since establishing criteria involves many value judgments that neither the human judge nor any AI-based system explicitly discusses (see also 4.3).

³ I am thankful to an anonymous reviewer for pointing this difference out.

4 Risks

There are many different AI-based systems and different use cases. Nevertheless, it is crucial to develop good risk assessment methods for each. For instance, the accuracy of automated vehicles can be compared with that of human drivers as a benchmark. Furthermore, such vehicles enable other strategies to reduce accident risk through interconnection. Furthermore, companies align their safety measures with existing ISO standards to minimize risk. In such standards, misuse scenarios must be envisioned to integrate countermeasures into the safety architecture. However, open-use systems, such as chatbots, that can be prompted for various use cases, such as a therapist, partner, friend, assistant, etc., are much more dangerous because companies cannot foresee all possible misuses or just test them for safety.

However, with regard to AI-based systems, it is not solely about their performance accuracy, as companies sometimes pretend. The probabilistic rationality of such systems must be made testable to achieve systems that function reliably and align with ethical values; thus, AI-based systems must be grounded in technical rationality. Behavioral experiments, for instance, could enable the development of AI therapists that are reliable and safe, and that reduce harmful consequences. Hubig distinguished between hypothetical and real risks: the former can be assessed through simulations, experiments, cross-sectional studies, etc., and the latter through calculations [27]. They can be mapped onto instrumental and technical rationality.⁴ I propose adding “predictive risks,” where the main focus on accuracy and optimization veils the more pressing question of assessing the ethical adequacy of such systems.

4.1 Real risk

The objective risk is often estimated as the frequency within a defined domain (i.e., the probability of occurrence) [46], p. 11, [27], p. 80, [35].

$$\text{Realrisk (chance)} = \frac{\text{Damage (benefit)}}{n}$$

For instance, the probability of an unwanted event can be thus estimated in the following way:

$$\frac{2788}{2406465} \approx 0.0012$$

⁴ Hubig describes a third risk/possibility: meta-risks which pose challenges mainly for our understanding of us as humans, for instance via genetic engineering.

2788 people died in a car accident (damage), while there were 2 406 465 accidents (n). The probability of dying in a car accident in Germany is 0.0012 based on the data from 2022 (Statistisches [57]). This only shows the frequency of an event, but the “risk” of car accidents outside the city or on the highway is higher because the velocity is higher, leading to more severe accidents. Furthermore, in the formula above, the damage or harm is already implicitly defined as fatal accidents. However, the (amount of) damage (depending on how it is determined: euros, hours for lack of work) could be multiplied by the probability of its occurrence (here, 0.0012) to depict the real risk. Real risks are, in the end, calculable and comparable, since it is possible to calculate fatal accidents in airplanes and then give “objective” probabilities of occurrence for the safety of certain means of transportation. Multiplied by the damage, it can lead to monetary comparability. For instance, the occurrence of a work accident results in the loss of a certain number of work hours.

There are several limitations to such a conceptualization of risk when putting it into context and evaluating its perception.

- (1) Risk is connected to moral concepts and values. For instance, “unreasonable risk” is defined as “risk (3.128) judged to be unacceptable in a particular context according to valid societal moral concepts.” [33] (3.176) The perception of risk of nuclear energy as unacceptable came with catastrophes, such as in Fukushima, and led countries like Germany to abandon this form of energy.
- (2) Furthermore, risk must always be operationalized, i.e., the criteria must be defined to open the search space for indicators and observables to quantify the risk. However, the observables can also be qualitative, as stated in ISO standards, where they discuss the “acceptance criterion,” which represents “the absence of an unreasonable level of risk,” whereas the “acceptance criterion can be of qualitative as well as quantitative nature, e.g., physical parameters that define when a specific behavior is considered as hazardous behavior, the maximum number of incidents per hour, as low as reasonably practicable (ALARP).” For example, from “traffic statistics, a reasonable level of risk of one accident per X km is derived.” [34] (3.1).
- (3) The analysis and evaluation of risks depend on a mixture of benefits and harms. Calculating risks seems a rational and objective way to handle situations of uncertainty. However, subjective risk should not be dismissed as irrational, as it can also be operationalized to some degree [46], p. 12). Objective risks might be action-guiding for engineers when developing technologies. Nevertheless, users might perceive subjective risks as more salient in deciding whether to do something or use a certain technology. For instance, the gratification of traveling to a nice vacation spot is not as high as the fear of flying and does not trump that perceived risk for some people, although the objective risk of flying is much lower than driving by car.
- (4) Also, according to such standards, risk in traffic is a real risk that at least can be calculated because data has already been collected (like the data about accidents in traffic), but for self-driving cars, such data does not directly exist. However, the Department of Motor Vehicles in California has published disengagement reports [12, 56]. Nevertheless, the calculation depends on the availability of good data.

4.2 Hypothetical risk

A possibility *de dicto* or hypothetical possibility refers to the status of a propositional content: it is possible that p [30]. This can be modeled as a possible world in which it is possible that something is the case. There might be a world in which Einstein had a brother. This world is accessible to us. (For a more technical analysis, see Kripke [38].) Most of such logical modeling is, to a certain degree, intuitive, nevertheless, “the *definition domains* within which these possibilities can occur are *not completely known*.” They are “hypothetical” or “theoretical” possibilities [27], p. 77). Many other propositions are already implicitly acknowledged and understood in everyday examples, as above, to formulate such a possible world. We were introduced to linguistic and cultural practices to learn what a brother is, who Albert Einstein was, etc. It becomes much more complicated in the case of scientific models or simulations.

Hypothetical risks are objectively imponderable. Unlike real risks, they cannot be calculated using probability calculus. This does not imply that no calculations can be made; rather, they differ from estimating the frequency of an unwanted event. Hypothetical risks are based on simulations, thought experiments [27], pp. 84–88), studies, or experiments. For instance, in x years, the place y will be unaffected by earthquakes, sedimentary movements, etc., so it can serve as a final storage for nuclear waste. Of course, no complete list of conditions can be given to exclude any possible movement. Simulations can be counterfactually robust to a certain degree, but there cannot be an exhaustive list of potential interferences to exclude every risk; uncertainty remains. Additionally, comparisons with similar practices,

technologies, and data collected for such technologies should be made to address residual risk.⁵

It can also be that there are either no parameters defined because one condition/possible interference is overlooked, taken as not relevant, or that no technical or experimental method is developed either to detect x or tests are only positive above a certain number of x in a sample (for instance, the number of viral RNA in samples [5]). Next to the scientific assessment, the risk assessment plays an important role. For instance, defenders of homeopathy could argue that the active ingredient is below the limit of detection. However, the perceived benefit may be overestimated through an anecdotal placebo effect. On the other hand, the risk perception of electric smog may be overestimated because it is not visible and the use of electric devices is ubiquitous, despite scientific evidence that it is not related to health problems.

Furthermore, it must also be distinguished between the opportunity to conduct experiments that establish causal links and observational studies that can only demonstrate a correlation. For instance, it is highly difficult to assess the risk of social media use and its potential linkage to mental health problems. Cross-sectional studies might find a correlation between them and suggest that the increase in mental health problems is due to social media use [59], whilst longitudinal studies might even suggest the inverse effect, i.e., depressive symptoms might lead to more social media use (among adolescent girls) [22]. Other studies have found that there might be a correlation between posting on social media (active use) and mental health problems, but not a correlation between screen time and mental health problems [65]. The concepts must be operationalized adequately, i.e., what social media use means, but certainly also what mental health problems are. Furthermore, other interfering factors must be excluded, such as increased economic instability, etc. In modeling such connections, value judgments are already included, particularly when technological systems and their use are observed with respect to values such as health. The problematic issue of adequately assessing such risks deepens when AI-based technologies are used to predict human behavior.

⁵ This is suggested from the EU for automated driving: “The manufacturer shall define the acceptance criteria from which the validation targets of the ADS are derived to evaluate the residual risk for the ODD taking into account, where available, existing accident data (1), data on performances from competently and carefully driven manual vehicles and technology state-of-the-art.

The manufacturer shall have processes to manage the safety and continued compliance of the ADS over lifetime (wear and tear of components especially for sensors, new traffic scenarios, etc.)” (EU Regulation 2022/1426, Annex II Performances, Article 7).

4.3 Predictive risk: accuracy and engagement vs. adequacy

Predictive risk relies on probabilistic rationality. For instance, people can be placed in certain risk categories based on the likelihood of repaying their loans or committing a crime. Whilst test scenarios in experimental settings can, to a certain extent, establish causal relationships, probabilistic rationality shows only correlations, where performance accuracy can be improved. In the case of the COMPAS tool to assess the risk of recidivism, a main point of critique is that it discriminates against minorities. This is due to a different conception of fairness used to assess it. On the one hand, it seems that the tool accurately predicts recidivism among black and white individuals, even though the performance is not that good for both groups. At least, this is what scientists from Northpointe, Inc., the company that developed COMPAS, presented. “Black defendants who were predicted to reoffend actually did recidivate at a slightly higher rate than their white counterparts (63% as against 59%).”⁶ Those results should provide “evidence of predictive parity” among the target population [39]. On the other hand, many more black persons are falsely classified in the high-risk category, and many more white persons are falsely classified in the low-risk category. The number of black defendants who were falsely classified as high risk (false positives) was “much higher than the corresponding percentage of whites (45% as opposed to 23%).” In opposition, white defendants were more often mistakenly classified as low risk (false negatives). “White reoffenders were mistakenly labeled as low risk almost twice as much as black ones (48% as against 28%)” [39] (see also Larson et al. [40]). A higher rate of false negatives is much more concerning since they pose a risk to public safety.⁷ Of course, mistakenly misclassifying someone as high-risk is not acceptable and tragic for the individual.

⁶ “Similarly, white defendants who were predicted not to recidivate did not reoffend at a slightly higher rate than black defendants (71% as against 65%)” [39].

⁷ Public safety is only one value that is at stake. The autonomy of the classified persons is also at stake if they cannot assess or contest the decisions that are made about them on certain rational grounds. This might also be linked with how the risks are communicated. The false-negative and false-positive rates might be difficult to understand; a more simplified version might be needed, as has been done in other contexts by, for instance, “communicating information using natural frequencies instead of probabilities” [26],[66]. Balancing values must be open to public discourse. This presupposes that the risks are adequately communicated and that users and citizens can set themselves in relation to the values, i.e., make autonomous judgments. (I am thankful to an anonymous reviewer for pointing out this tension).

Hellman calls the former conception of fairness “predictive parity” and the latter “equalized odds.” She states that an “algorithmic tool cannot satisfy both predictive parity and equalized odds when base rates are different unless the tool is perfectly accurate.” [23] However, both fairness conceptions fall under the paradigm of “predictive risks.” In a certain sense, predictions must be made to protect the general public and to institute public safety measures by imposing longer sentences. However, predictive power that aims to accurately predict individuals’ future behavior must be based on transparent criteria for categorization. Such criteria can be challenged and discussed in public forums on their adequacy. One criterion is stability and coherence in decision-making (finishing school, keeping a job, etc.); another criterion might be related to personality traits (short-temperament, violent behavior, etc.). Those are valid indicators and also part of the survey questions, which serve as input data for the COMPAS algorithm. However, (1) they are backward-looking and do not take into account future stability, nor how to provide social or institutional structures for convicts to prevent recidivism. Furthermore, (2) scientific studies are needed to establish connections between certain factors (such as not finishing school) and criminal activities. Assuming, for instance, that tax evaders probably have a higher education and might also reoffend, it might be a bad indicator for certain groups (also assuming that they might be mostly white) and thus lead to misclassification among white convicts (false negatives). Predictive power is seductive and an easy way out, since it makes social issues quantifiable and veils their problematic nature by providing supposedly “accuracy” for predictions.

Another example is ideological polarization through social media and its algorithms, which optimize for engagement rather than truth or accuracy. Dorsch et al. argue that not the so-called “filter bubbles” directly reinforce ideological polarization. Instead, “engagement architectures, optimized precisely for fluent processing,” as in social media, inflate “confidence in the user’s own epistemic stance.” They suggest introducing “epistemic friction” to avoid ideological dazzlement. This friction might “promote reflective evaluation and deliberative reasoning” by “asking users to justify their beliefs or explain their reasoning, as this too would introduce disfluency, since the act of epistemic justification is cognitively demanding.” [13] Such engagement architectures rely on predictive rationality to optimize, based on collected data, what might be of interest for the user [45], p. 94/95).

5 Concluding remarks—autonomy and power

The exercise of power is sometimes subtle and masked, stemming from the architectural design of decision-making and different forms of rationality. It should, however, not be seen as something purely negative if adequate criteria are developed to ensure ethical alignment. This is achieved through participatory means, where citizens and users are included in the development of technologies and their autonomy is respected. This contributes to a kind of legitimacy by incorporating reflected value decisions that are accessible to users and open to critique. Every complex social system has a normative framework in which norms and rules are explicitly written down, and forms of protest are permitted. Socio-technical systems should be developed to ensure that they do not “undermine the autonomy of decisions” [29], p. 142). This implies that protest must be possible without manipulating, coercing, or punishing. The recognition of autonomy is logically prior to any other value; otherwise, ideological dazzlement may arise, as in algorithms that optimize for engagement. Instrumentalization of individuals through control, whether by surveillance or other technological means, conditions behavior. Regulating behavior based on predictive power is an encroachment on autonomy, which is often a more subtle influence on human behavior without having explicit values and criteria developed. Engagement architectures are designed to hinder epistemic friction and deliberative processes. Hence, the implementation of technologies must be carried out in a participatory manner, through deliberative mini-publics, parliamentary legislation, or governance measures, such as standardization programs (ISO, IEEE, etc.), to obtain legitimacy and to make the ethical acceptability of such systems accessible. For this purpose, norms must be explicit and transparent. Additionally, it must be possible to protest, change norms, retract technological developments, etc.; otherwise, the autonomy of users and citizens is not respected, and they are overpowered in favor of questionable measures to minimize risks.

Author contributions Single author, F.R.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate

if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acemoglu, D., Johnson, D.: Power and progress: our thousand-year struggle over technology and prosperity. Basic Books UK, London (2023)
- Amann, E.: Entscheidungstheorie: Individuelle, strategische und kollektive Entscheidungen. s.n., Wiesbaden (2019)
- Anderson, J.: Understanding Cultural Geography: Places and Traces. Routledge, New York (2010)
- Aristoteles, 1995. *Nikomachische Ethik*. nach der Übersetzung von Eugen Rolfes, bearbeitet von Günther Bien ed. Hamburg: s.n.
- Arnaut, R., et al.: The limit of detection matters: the case for benchmarking severe acute respiratory syndrome coronavirus 2 testing. Clin. Infect. Dis. (2021). <https://doi.org/10.1093/cid/ciaa1382>
- Ashby, W.R.: Einführung in die Kybernetik. Suhrkamp, Frankfurt am Main (1974)
- Birhane, A. et al., 2022. Power to the people? Opportunities and challenges for participatory AI. Proceedings of the 2nd ACM Conference on Equity and Access in Algorithm, Mechanisms, and Optimization.
- Breiman, L.: Statistical modeling: the two cultures. Stat. Sci. **16**(3), 199–231 (2001)
- Bungartz, H.-J., Zimmer, S., Buchholz, M., Plfüger, D.: Modellbildung und Simulation: Eine anwendungsorientierte Einführung. Springer, Heidelberg (2009)
- Cooper, A. & Abrams, E., 2021. Emergent unfairness in algorithmic fairness-accuracy trade-off research. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 46–54.
- Cox, D.: Comment on: statistical modeling: the two cultures. Stat. Sci. **16**(3), 216–218 (2001)
- DMV State of California, 2023. *Disengagement Reports*. [Online] Available at: <https://www.dmv.ca.gov/portal/es/vehicle-industry-services/autonomous-vehicles/disengagement-reports/>
- Dorsch, J., Coeckelbergh, M., Vierkant, T.: Does AI undermine democratic deliberation? How engagement algorithms hijack epistemic agency. Synthese (2026). <https://doi.org/10.1007/s11229-026-05820-6>
- Floridi, L., Cowls, J., King, T., Taddeo, M.: How to design AI for social good: seven essential factors. Sci. Eng. Ethics **26**(3), 1771–1796 (2020)
- Foucault, M.: Discipline and Punish: The Birth of the Prison. Vintage Books, New York (1979)
- Foucault, M.: The subject and power. Crit. Inq. **8**(4), 777–795 (1982)
- Foucault, M.: Das Spiel des Michel Foucault. In: Schriften in 4 Bänden (Dits et Ecrits), pp. 391–429. Suhrkamp, Frankfurt am Main (2003)
- Friedman, B., Kahn, P., Jr., Borning, A.: Value sensitive design and information systems. In: Kimma, K., Tavani, H. (eds.) The Handbook of Information and Computer Ethics, pp. 69–101. Wiley, New Jersey (2008)
- Grunwald, A.: Technology assessment or ethics of technology? Ethic. Perspect. **6**(2), 170–182 (1999)
- Grunwald, A., Julliard, Y.: Technik als Reflexionsbegriff: Überlegungen zur semantischen Struktur des Redens über Technik. Philosophia naturalis, Issue **42**, 127–157 (2005)
- Han, B.-C.: Psychopolitics: Neoliberalism and New Technologies of Power. Verso, London/New York (2017)
- Heffer, T., et al.: The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: an empirical reply to. Clin. Psychol. Sci. **7**(3), 462–470 (2019)
- Hellman, D., 2025. Algorithmic Fairness. The Stanford Encyclopedia of Philosophy.
- Himmelreich, J.: Against “democratizing AI.” AI Soc. (2023). <https://doi.org/10.1007/s00146-021-01357-z>
- Hoadley, B.: Comment on: statistical modeling: the two cultures. Stat. Sci. **16**(3), 220–224 (2001)
- Hoffrage, U., Gigerenzer, G.: Using natural frequencies to improve diagnostic inferences. Acad. Med. **73**(5), 538–540 (1998)
- Hubig, C.: Technik- und Wissenschaftsethik: Ein Leitfadens, 2 ed Springer, Heidelberg/Berlin (1995)
- Hubig, C.: Die Kunst des Möglichen I: Technikphilosophie als Reflexion der Medialität. s.n., Bielefeld (2006)
- Hubig, C.: Die Kunst des Möglichen II: Ethik der Technik als provisorische Moral. Transcript, Bielefeld (2007)
- Hubig, C., 2010. Möglichkeit. *Enzyklopädie Philosophie*, Volume 2, I - P, pp. 1642–1649.
- Hubig, C., 2011. „Natur“ und „Kultur“: Von Inbegriffen zu Reflexionsbegriffen. *Zeitschrift für Kulturphilosophie*, Volume 5,1, pp. 97–119.
- Hubig, C.: Die Kunst des Möglichen III: Macht der Technik. Transcript, Bielefeld (2015)
- ISO, 2018. *ISO 26262-1:2018 Road vehicles - Functional safety*, s.l.: s.n.
- ISO, 2022. *ISO 21448:2022 Road vehicles - Safety of the intended functionality*, s.l.: s.n.
- Kaplan, S., Garrick, B.J.: On the quantitative definition of risk. Risk Anal. (1981). <https://doi.org/10.1111/j.1539-6924.1981.tb01350.x>
- Klenk, M.: Ethics of generative AI and manipulation: a design-oriented research agenda. Ethics Inf. Technol. (2024). <https://doi.org/10.1007/s10676-024-09745-x>
- Kosinski, M., Wang, Y.: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. J. Pers. Soc. Psychol. **114**(2), 246–257 (2018)
- Kripke, S.: Semantical analysis of modal logic I normal modal propositional calculi. Math. Log. Q. **9**, 67–96 (1963)
- Lagioia, F., Rovatti, R., Sartor, G.: Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. AI Soc. **38**, 459–478 (2023)
- Larson, J., Mattu, S., Kirchner, L. & Angwin, J., 2016. How We Analyzed the COMPAS Recidivism Algorithm. [Online] Available at: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> [Accessed 5 February 2026].
- Libai, B., Kaplan, A., Kötterheinrich, K., Kroll, E.B.: Brave new world? On AI and the management of customer relationships. J. Interact. Mark. **51**(1), 44–56 (2020)
- Lin, T.: Democratizing AI” and the concern of algorithmic injustice. Philos. Technol. (2024). <https://doi.org/10.1007/s13347-024-00792-2>
- Macnish, K., van der Ham, J.: Ethical Approaches to Cybersecurity. In: Véliz, C. (ed.) The Oxford Handbook of Digital Ethics. Oxford University Press, s.l. (2022)
- Merchant, R., et al.: Evaluating the predictability of medical conditions from social media posts. PLoS ONE (2019). <https://doi.org/10.1371/journal.pone.0215476>

45. Mühlhoff, R.: The Ethics of AI: Power, Critique, Responsibility. Bristol University Press, Bristol (2025)
46. Nida-Rümelin, J., Rath, B., Schulenberg, J.: Risikoethik. s.n., Göttingen (2012)
47. Noorman, M., Swierstra, T.: Democratizing AI from a sociotechnical perspective. *Minds Mach.* **33**(4), 563–586 (2023)
48. Peterson, M.: An Introduction to Decision Theory, 2 ed Cambridge University Press, Cambridge (2017)
49. Richter, F.: AI ethics as a form of research ethics. In: What AI Can Do, pp. 5–22. Chapman and Hall/CRC, Boca Raton/London/New York (2023)
50. Richter, F.: Justice and Smart Societies: Conceptual Foundations, Limitations, and Conditions of Algorithmizing Fairness. In: C. Didier, et al. eds. Engineering and Value Change. Philosophy of Engineering. Cham: Springer (2025)
51. Richter, F.: From human-system interaction to human-system co-action and back: ethical assessment of generative AI and mutual theory of mind. *AI & Ethics.* **5**, 19–28 (2025)
52. Richter, F., Krügel, S., Uhl, M.: Users prefer generative AI to reflect real over idealized gender proportions. *Sci. Rep.* (2026). <https://doi.org/10.1038/s41598-026-66305-7>
53. Ropohl, G.: Technikbegriffe zwischen Äquivokation und Reflexion. In: Banse, G., Grunwald, A. (eds.) Technik und Kultur: Bedingungs- und Beeinflussungsverhältnisse, pp. 41–54. s.n., Karlsruhe (2010)
54. Rosenberger, R., Verbeek, P.-P.: A postphenomenological field guide. In: Rosenberger, R., Verbeek, P. (eds.) Postphenomenological Investigations: Essays on Human-Technology Relations, pp. 9–41. s.n., New York/London (2015)
55. Sich, P.: Foucault: Eine Einführung. Reclam, Ditzingen (2018)
56. Sinha, A., et al.: Crash and disengagement data of autonomous vehicles on public roads in California. *Sci. Data* (2021). <https://doi.org/10.1038/s41597-021-01083-7>
57. Statistisches Bundesamt, 2023. *Gesellschaft und Umwelt: Verkehrsunfälle*. [Online] Available at: https://www.destatis.de/DE/The men/Gesellschaft-Umwelt/Verkehrsunfaelle/_inhalt.html#238556
58. Thaler, R.H., Sunstein, C.R.: Nudge: The Final Edition. Penguin Books, s.l. (2021)
59. Twenge, J., Joiner, T., Rogers, M., Martin, G.: Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. *Clin. Psychol. Sci.* **6**(1), 3–17 (2018)
60. Umbrello, S., van de Poel, I.: Mapping value sensitive design onto AI for social good principles. *AI & Ethics* **1**(3), 283–296 (2021)
61. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. *Minds Mach.* **30**, 385–409 (2020)
62. Verbeek, P.-P.: Beyond interaction: a short introduction to mediation theory. *Interactions* **22**(3), 26–31 (2015)
63. Weber, M.: Wirtschaft und Gesellschaft: Grundriß der verstehenden Soziologie (1. Halbband). s.n., Tübingen (1921)
64. WHO, 1946/48. *Constitution of the World Health Organization*. [Online] Available at: https://apps.who.int/gb/bd/pdf_files/BD_49th-en.pdf [Accessed 23 9 2026].
65. Yu, Y., Dykxhoorn, J., Plackett, R.: The impact of different types of social media use on the mental health of uk adults: longitudinal observational study. *J. Med. Internet Res.* **26**, e56950 (2024)
66. Zhu, L., Gigerenzer, G.: Children can solve Bayesian problems: the role of representation in mental computation. *Cognition* **98**(3), 287–308 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.