CHANGEK

CHALLENGES AND INNOVATIVE CHANGES IN RESEARCH ETHICS REVIEWS

Review Guidelines on Artificial Intelligence



Author: Alexander Bagattini (ITZ)

Reviewers: Alexandra Zorila (UB), Constantin Vica (UB); Pascal Borry (KUL).

Document Revision History *(including peer reviewing & quality control)*

Version	Date	Changes	Contributor(s)
V1.0	10.04.26	General structure deliverable + electronic informed consent guideline added	Alexander Bagattini
V2.0	01.06.26	Integration of feedback received from CAHNGER partners on AI guideline	*Reviewers:* Alexandra Zorila, Constantin Vica, Pascal Borry. *Adaptations*: Alexander Bagattini

Pending approval from the EC.

Table of Contents

Guideline 3: Artificial intelligence 5
1. Introduction 5
1.1 The REC's role in the responsible governance of AI research 6
2. Framing the ethics review context for artificial intelligence 7
2.1 What do we mean when we speak of AI? 7
2.2 How AI research has shaped the world and may shape the future 8
2.3 Research on AI and research with AI 8
2.4 Why AI poses distinctive problems for ethics review 9
2.5 Legal and governance context 9
3. Research on AI: Non-Discrimination and Algorithmic Fairness 10
3.1 Sources of bias across the AI lifecycle 10
3.2 Fairness metrics and normative trade-offs 11
3.3 Individual and group harms 11
3.4 Human rights, vulnerable groups, and shared values 12
3.5 Defining the scope: legal issues versus ethical issues 12
3.6 Recommendations 13
4. Research on AI: Autonomy, agency and control 13
4.1 Meaningful human control 13
4.2 Privacy and data governance 14
4.3 Manipulation, dependency, and behavioural steering 14
4.4 Democracy, equity, and power asymmetries 15
4.5 Recommendations 15
5. Research on AI: Safety, Security and Well-Being 15
5.1 Safety and robustness 16
5.2 Security 16
5.3 Explainability and the need to understand AI 16
5.4 Dual use and misuse 17
5.5 Beneficence and non-maleficence 17
5.6 Recommendations 17

6. Research with AI: Research Integrity, AI, and the Shapes of Contemporary Science 18
6.1 Research integrity and AI 18
6.2 Research integrity and epistemic principles 19
6.3 Authorship, responsibility, and accountability 19
6.4 Deskilling and upskilling 20
6.5 Missing understanding and black boxes in science 20
6.6 Recommendations 20
7. Pulling the ethics review together 21
7.1 Key questions for RECs 21
7.2 What researchers should include in the protocol 22
7.3 Red flags 22
7.4 Cross-cutting recommendations 22
8. References and further resources 23

List of Tables

(If any)

List of Terms & Abbreviations

Abbreviation	Definition
AI	Artificial intelligence
GenAI	Generative artificial intelligence
GPAI	General-purpose AI
HLEG	High-Level Expert Group on AI
REC	Research ethics committee
RMF	Risk management framework

1. Introduction

Artificial intelligence (AI) can be understood, for the purposes of this guidance document, as machine-based systems that are designed to operate with varying degrees of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, decisions, or content that influence physical or virtual environments (European Parliament and Council 2024; OECD 2024). This broad understanding is useful for research ethics review, because it captures not only prominent contemporary forms of AI, such as machine learning systems and generative AI models, but also a wider range of computational systems that may shape research practices, participant experiences, institutional decisions, and socially significant outcomes. At the same time, major international governance frameworks have emphasized that AI should be developed and used in ways that remain human-centered, rights-respecting, and attentive to fairness, transparency, accountability, safety, and societal well-being (UNESCO 2021; OECD 2024).

AI has become increasingly important in contemporary research both as an object of investigation and as a tool used in the production of knowledge. Researchers may develop, train, test, validate, or deploy AI systems in areas such as healthcare, education, public administration, and security, to name just a few of them. They may also use AI within the research process itself, for example to classify and analyze data, support prediction and decision-making, generate text or images, to assist in literature review and scientific writing, or to generate reasonable hypotheses. These developments create important opportunities for innovation and scientific progress. However, they also raise distinctive ethical concerns, including the risk of bias and discrimination, diminished transparency, weakened human oversight, privacy and data governance challenges, safety and security risks, and questions concerning accountability and responsibility. In addition, the use of AI in research may affect the reliability, interpretability, and integrity of scientific knowledge itself.

For Research Ethics Committees (RECs), these developments create a dual challenge. On the one hand, RECs must assess research *on* AI, including studies in which AI systems are developed, evaluated, or introduced into socially relevant settings. On the other hand, they must also assess research *with* AI, that is, research in which AI functions as a methodological, analytic, or epistemic tool. In both cases, the task of ethics review extends beyond narrow legal compliance. It requires examining whether the proposed training, design, use, and governance of AI are ethically justifiable in light of the potential effects on research participants, affected groups and communities, researchers, institutions, and society more broadly. This includes attention not only to individual harms, but also to structural unfairness, power asymmetries, epistemic risks, and the conditions under which meaningful human oversight can be sustained.

This guidance document is intended to support RECs, research ethics officers, administrators, and researchers in identifying and evaluating these issues in a structured and practically relevant way. It does not aim to replace existing legal or regulatory frameworks, but rather to complement them by highlighting the core ethical questions that arise when AI is developed or used in research. It also builds on current European discussions about how ethics review should adapt to AI-rich research environments. Work in the iRECS project has underlined the need for stronger AI-related expertise

and more continuous forms of ethics review, while AIOLIA has shown how general principles such as fairness, autonomy, safety, and accountability can be translated into operational and reviewable components (Aucouturier and Grinbaum et al. 2023; Aucouturier and Grinbaum 2024; Teo, Kyosovska, and Armengol 2025; Bayerl et al. 2026). The sections that follow therefore distinguish between key areas of concern that are particularly important in the ethics review of AI-related research: non-discrimination and algorithmic fairness, autonomy and meaningful human control, safety, security and well-being, and the implications of AI for research integrity and the changing shape of contemporary science.

1.1 The REC's role in the responsible governance of AI research

RECs are not substitutes for technical validators, data protection officers, cybersecurity specialists, or competent regulatory authorities. Their distinctive contribution lies elsewhere: ethically acceptable AI in research cannot be secured by legal compliance or benchmark performance alone. The REC is often the place within an institution where scientific validity, proportionality, participant protection, social consequences, and the conditions of responsible oversight are assessed together. That integrative role matters especially for AI because many of the most important ethical problems appear at the interfaces between domains: between data governance and fairness, between explainability and safety, between human oversight and responsibility, or between efficiency gains and research integrity.

In practice, this means that AI-related review asks RECs to address questions that may once have sat at the margins of ethics review. Is the training data appropriate for the population in which the system will be used? What forms of automation bias or over-reliance are foreseeable? Who is responsible for monitoring drift, failure, or harmful downstream effects after approval? Does the protocol contain a credible plan for human oversight, or only a symbolic reference to 'the human in the loop'? Are there communities that may absorb the burdens of error, exclusion, or surveillance without sharing in the benefits of the research? These are not merely technical or legal questions. They are normative questions about acceptable risk, distribution of burdens, respect for persons, and institutional accountability.

For these reasons, this guidance document treats RECs as bodies that should be able to identify when AI-specific expertise is required, request additional documentation, require proportionate safeguards, and insist on review pathways that extend beyond the first approval decision for dynamic or high-risk systems. This is consistent with recent recommendations from iRECS, which emphasize that AI-related research review requires stronger scientific expertise within or around RECs, and with the broader move in Europe toward more accountable and risk-sensitive AI governance (Aucouturier and Grinbaum et al. 2023; European Commission 2025a).

2. Framing the ethics review context for artificial intelligence

2.1 What do we mean when we speak of AI?

For the purposes of ethics review, "AI" should not be reduced to a single technical category. In research settings the term may refer to supervised, unsupervised, or reinforcement learning; to

computer vision, natural language processing, recommender systems, large language models, world models, multimodal foundation models, predictive analytics, or generative systems; and to tools that operate as classifiers, ranking systems, decision-support systems, conversational agents, content generators, or autonomous optimization systems. The review question is therefore not simply whether a project "uses AI", but what kind of AI it uses, for what task, on what data, with what degree of adaptivity, opacity, autonomy, and specific consequence.

This matters because ethically relevant features vary substantially between AI systems. A static model validated on a closed dataset is different from a continuously updated model embedded in an institutional workflow. A generative model used for brainstorming research questions is different from an AI tool that supports triage decisions in a clinical trial, automates welfare eligibility assessments, or evaluates student performance. RECs therefore need conceptually precise descriptions of the system architecture, intended use, development stage, and interaction context. The AI Act itself reflects this by combining a risk-based framework with categories such as prohibited practices, high-risk systems, general-purpose AI models, and additional duties for specific uses (European Parliament and Council 2024).

A practical definition for REC work is that AI refers to computational systems that infer from data and produce outputs such as predictions, recommendations, classifications, rankings, or generated content that can shape research processes, participants' experiences, or socially relevant decisions. This definition keeps attention focused on probabilistic inference, output, context, and effect rather than on branding or novelty claims alone.

2.2 How AI research has shaped the world and may shape the future

AI research has already changed everyday life and institutional practice. It has altered how text, images, and data are processed; accelerated some forms of pattern recognition; and enabled new forms of automation in health care, education, policing, finance, social media governance, and science itself. In research contexts, AI can help analyze large datasets, detect patterns in imaging, draft code, cluster literature, support translation, and generate hypotheses. It may thereby improve efficiency and, in some domains, enable forms of inquiry that would otherwise be infeasible. At the same time, the same features that make AI attractive (scale, speed, adaptability, and data hunger) also amplify familiar ethical concerns: discrimination can occur at scale; security failures may propagate quickly; manipulation can become personalized; and flawed outputs may acquire unwarranted authority because they appear technically sophisticated (UNESCO 2021; OECD 2024).

The likely future impact of AI on research is not only methodological but also institutional and epistemic. AI systems may reorder who can do research, what counts as expertise, how results are justified, and which infrastructures become indispensable. Proprietary foundation models, cloud dependency, unequal compute access, and opaque model supply chains all have implications for scientific autonomy and fairness. The European Commission's work on generative AI in research rightly stresses that AI in science is not just a question of tool adoption but of research culture, integrity, accountability, and the public interest (European Commission 2025a).

For RECs, the future-oriented dimension is important because some AI harms are anticipatory. They involve not only actual current harms, but plausible pathways toward dependency, deskilling, lock-in, or normalization of ethically questionable practices. A good review therefore asks not only whether a system is currently lawful or beneficial, but whether the proposed research contributes to a trajectory that is ethically defensible.

2.3 Research on AI and research with AI

A distinction that is central both to the CHANGER MOOC and to this guidance document is the distinction between research on AI and research with AI. Research on AI includes the design, development, training, tuning, validation, testing, deployment, and evaluation of AI systems themselves. Here the AI system is the object of research, and the main ethical questions concern how it is built and introduced. Research with AI, by contrast, refers to the use of AI as a tool in the conduct of research: for instance, for literature review, coding, text production, image generation, data analysis, participant interaction, or evaluation workflows. Here the AI system is part of the research process, and the ethical questions concern integrity, transparency, authorship, competence, reproducibility, and the reliability of the knowledge produced. Importantly, different ethical issues emerge when doing research with or on AI.

The distinction is analytically useful because it helps RECs identify which ethical issues are primary. If a study develops a predictive model for clinical deterioration, fairness, safety, oversight, and validation are likely to be central. If a research team uses generative AI to summarize literature, draft sections of a manuscript, or assist in qualitative coding, then disclosure, verification, accountability, epistemic quality, and the risk of deskilling become more salient. Many projects involve both dimensions at once. A study may, for example, develop a generative tool for research administration while simultaneously relying on AI-generated content during protocol preparation. RECs should therefore permit mixed classifications rather than forcing a project into only one category.

This distinction also clarifies why AI ethics review cannot be reduced to product safety or data protection review. Research with AI may raise serious integrity issues even where no participant-facing harm is obvious, while research on AI may raise participant and societal risks even where the project's publication practices are beyond any doubt.

2.4 Why AI poses distinctive problems for ethics review

AI presents distinctive challenges for ethics review because ethically relevant properties may be distributed across datasets, model architectures, prompts, interfaces, institutional workflows, users, and post-deployment conditions. Risk is therefore not always legible at the point of protocol submission. Data provenance may be uncertain. Model behaviour may be unstable across contexts. Feedback loops may emerge only after deployment. Human oversight may degrade over time as users become more reliant on automation. And harmful effects may be borne not only by identifiable participants, but also by indirectly affected groups who are absent from the study design.

Another challenge is that AI research often crosses conventional governance silos. A single project may raise questions about human subjects' protection, personal data processing, copyright, information security, clinical safety, discrimination, platform dependency, and scientific integrity.

Existing institutional structures sometimes fragment these questions rather than integrating them. The result can be a form of ethical under-review in which each office assumes that another office is addressing the core risk. RECs can help counter this fragmentation by requiring protocols to clarify what has been reviewed elsewhere and what remains ethically open.

Finally, AI can make reviewing itself more difficult. Technical complexity, rapid change, uncertain terminology, and commercially mediated infrastructures can make it hard for non-specialists to interrogate system claims. iRECS has shown that this problem is already felt in REC practice and that relying on ordinary review procedures without strengthening expertise leaves committees poorly equipped to scrutinize AI-related research adequately (Aucouturier and Grinbaum 2024).

2.5 Legal and governance context

The ethical review of AI research takes place within a rapidly evolving governance landscape. In the European Union, the AI Act entered into force on 1 August 2024 and becomes fully applicable from 2 August 2026, with some obligations applying earlier or later depending on the category, including prohibited practices from February 2025, obligations for general-purpose AI from August 2025, and some high-risk product-related rules from August 2027 (European Commission 2025b; European Parliament and Council 2024). For RECs, the key point is not to act as AI Act enforcement bodies, but to recognize when a protocol touches categories of risk or governance addressed by the Act and to require researchers to explain how legal duties intersect with their ethical safeguards. (Schmidt, M. W., Bagattini, A., Pfister, J., Martin, D., & Hillerbrand, R. 2026)

The AI Act does not exhaust the relevant framework. Depending on the project, RECs must also be attentive to data protection rules, anti-discrimination law, medical device regulation, product safety, research integrity standards, cybersecurity requirements, sector-specific governance, and institutional policies. At the level of principles and soft law, especially influential sources include the UNESCO Recommendation on the Ethics of Artificial Intelligence, the OECD AI Principles, the Ethics Guidelines for Trustworthy AI, ALTAI, NIST's AI Risk Management Framework and Generative AI Profile, and the European Research Area's living guidelines on the responsible use of generative AI in research (UNESCO 2021; OECD 2024; High-Level Expert Group on AI 2019, 2020; NIST 2023, 2024; European Commission 2025a).

The role of this guidance document is therefore complementary. Rather than restating high-level governance principles in the abstract, it translates them into REC-relevant questions about participant protection, fairness, control, safety, and integrity. In doing so, it also takes seriously a central lesson emphasized by AIOLIA: ethical guidance for AI is most useful when it moves beyond general commitments and is made operational through context-sensitive procedures, clearly assigned responsibilities, and practically usable indicators for evaluation and oversight. AIOLIA's work is particularly helpful in showing that principles such as fairness, human oversight, accountability, and transparency should not be treated as self-interpreting ideals, but as aims that must be specified in relation to the concrete research setting, the actors involved, the stage of the AI lifecycle, and the kinds of risks that may arise for individuals, groups, and institutions. For RECs, this means that ethical review should not stop at asking whether such principles are mentioned in a protocol, but should

examine how they are implemented, monitored, and justified in practice (Teo, Kyosovska, and Armengol 2025; Bayerl et al. 2026).

3. Research on AI: Non-Discrimination and Algorithmic Fairness

3.1 Sources of bias across the AI lifecycle

Fairness problems in AI rarely arise at a single point. They can emerge in data collection, annotation, model design, evaluation, deployment, and institutional uptake. Training data may underrepresent certain groups, overrepresent historical disadvantage, rely on poor proxies, or reflect discriminatory institutional practices. Labels may encode contested or biased judgments. Feature selection may privilege variables correlated with protected characteristics. Benchmark datasets may not represent the population in which the system will be used. User interfaces may encourage uncritical uptake of model outputs, thereby turning moderate statistical bias into severe social harm. A fair ethics review must therefore examine the entire lifecycle rather than asking only whether the final model performs equally well on average (Chen et al. 2023; Koçak et al. 2025).

This lifecycle perspective is especially important in research settings because methodological decisions are often presented as purely technical even when they are value-laden. Decisions about what outcome to predict, which data to exclude, how to define error, and what counts as “ground truth” all have ethical significance. In health-related AI, for example, unequal access to care can distort the data from which models learn. In social data research, historical policing or content moderation patterns may encode structural bias. RECs should therefore ask researchers to explain the social provenance of their data and the normative assumptions built into labels, categories, and success metrics.

Recent literature in medicine and healthcare illustrates how easily fairness failures can become clinically relevant. Reviews by Chen et al. and Koçak et al. show that data imbalance, spectrum bias, site-specific bias, and hidden confounding can systematically disadvantage certain populations, even where the model appears technically strong in aggregate evaluations. Similar concerns arise far beyond healthcare, including education, employment, and public administration, where population shifts and proxy labels may quietly magnify existing inequalities (Chen et al. 2023; Koçak et al. 2025).

A more explicitly research-ethics-oriented literature sharpens this point further. It has been argued that algorithmic bias should not be treated as an isolated technical defect, but as a socio-technical phenomenon that can emerge from problem formulation, proxy selection, temporal feedback loops, and institutional design choices. Their analysis is especially useful for RECs because it shows why fairness review must reach beyond dataset composition to the normative assumptions built into prediction targets, time horizons, and governance arrangements (Bagattini and Shao 2025).

3.2 Fairness metrics and normative trade-offs

There is no single fairness metric that solves all ethical concerns. Statistical parity, equalized odds, predictive parity, calibration, and individual fairness capture different ideas of what it means to treat persons or groups fairly. In some settings these metrics are mutually incompatible. This means that fairness is not a purely technical optimization problem, but a normative choice about which forms of

inequality matter most in the context at hand. For RECs, the practical implication is that researchers should not be allowed to claim that a system is "fair" without specifying the fairness objective, why it was selected, what trade-offs it creates, and who bears the residual burdens. In high-stakes settings, committees should expect justification of the chosen metric in relation to the values and harms specific to the use case. A diagnostic tool, for example, may warrant special concern for false negatives in underdiagnosed populations. A triage or ranking tool may require scrutiny of error distribution across protected and vulnerable groups. Where fairness metrics conflict, researchers should explain whose interests are prioritized and why.

This is another area where AIOLIA is instructive. Its work underscores that principles such as fairness and non-discrimination become reviewable only when broken down into practical components, responsibilities, and evidence. An ethics guideline that stops at the word "fairness" without demanding operationalization leaves RECs with little to assess (Bayerl et al. 2026).

3.3 Individual and group harms

Fairness failures can produce both individual and group harms. At the individual level, model error may result in exclusion, delayed care, misclassification, denial of opportunity, denial of access to resources, or inappropriate intervention. At the group level, repeated small errors can stigmatize communities, reinforce stereotypes, shift institutional resources away from already disadvantaged populations, or legitimize unequal treatment under the appearance of objectivity. Some harms are direct and immediate; others are cumulative, symbolic, or structural.

Importantly, group harms may occur even where no single participant can show a large measurable loss. This poses a challenge for standard REC reasoning, which often focuses on identifiable participant risk. In AI-related studies, the relevant ethical question may instead concern the distribution of burdens across social groups, including those not directly enrolled in the study. RECs should therefore ask, in a reasonable way, whether the protocol includes impact assessment for affected groups, not merely for recruited individuals.

Recent literature and patient-oriented reviews also suggest that fairness, transparency, and voice affect public trust. Patients and other data subjects are often less concerned with AI in the abstract than with whether it will treat them as substitutable data points, whether it will obscure responsibility, and whether errors will fall unevenly on those who already experience disadvantage (Bashkin 2025).

3.4 Human rights, vulnerable groups, and shared values

Fairness review should not be detached from broader human rights and democratic values. The most pressing question is often not whether a model is equally accurate across predefined groups, but whether the research itself is compatible with dignity, non-discrimination, inclusion, due process, and contestability. This becomes particularly important for projects involving children, persons with disabilities, migrants, racialized minorities, linguistically marginalized groups, or socioeconomically disadvantaged communities. In these settings, a technically well-performing system may still be ethically problematic if it is imposed without meaningful consultation, does not permit challenge, or amplifies pre-existing asymmetries in voice and power (UNESCO 2021; OECD 2024).

The language of shared values is helpful here because it links fairness to legitimacy. RECs should ask not only whether a system avoids measurable bias, but whether the protocol is responsive to the values and expectations of the populations whose lives it may shape. Co-design, community consultation, and participatory methods may therefore be ethically important, especially where the social meaning of error or classification is contested. The iRECS and AIOLIA projects both reinforce the importance of context-sensitive ethical interpretation rather than generic principle recital (Aucouturier and Grinbaum et al. 2023; Teo, Kyosovska, and Armengol 2025).

3.5 Defining the scope: legal issues versus ethical issues

Many fairness-related concerns also have legal dimensions. Anti-discrimination law, equality duties, consumer protection, and sectoral regulation may all apply. But legal compliance is not identical to ethical acceptability. First, the law may lag behind technical practice or may not address cumulative and group-level harms adequately. Second, ethics review often asks questions about proportionality, respect, trustworthiness, and accountability that go beyond narrow compliance. Third, some legally permitted forms of optimization may still be ethically questionable because they exploit vulnerability, rely on opaque proxies, or offload risk onto groups with less institutional power.

RECs should therefore resist two errors: treating fairness as purely legal and hence somebody else's responsibility, or treating fairness as purely technical and therefore beyond ethical review. The right approach is integrative. Researchers should identify applicable legal requirements, but they should also articulate the ethical rationale for their fairness strategy, including the limits of that strategy.

3.6 Recommendations

- Require researchers to describe the full AI lifecycle relevant to the study, including data provenance, labelling, model development, evaluation, deployment context, and monitoring plans.
- Require explicit justification of the fairness objective used in the project. Claims that a system is "fair" should specify the metric, the reason for selecting it, and the trade-offs created.
- Ask for subgroup performance evidence where ethically relevant, especially in studies involving protected or vulnerable populations.
- Require discussion of individual and group harms, including reasonably foreseeable indirect harms to non-participating communities, while distinguishing them from remote or purely speculative risks. Where labels or proxies are socially contested, require researchers to explain their normative assumptions and the possible implications of those assumptions.
- Where risks of exclusion, stigmatization, or structural disadvantage are foreseeable, require participatory input, community consultation, or other context-sensitive forms of stakeholder engagement.
- Do not treat legal compliance as sufficient. Ask whether the fairness strategy is ethically proportionate to the stakes of the use case.
- Ensure that AI-related proposals are reviewed with access to appropriate ethics and domain expertise, either within the REC or through external consultation.4. Research on AI: Autonomy, agency and control.

4. Research on AI: Autonomy, agency and control

4.1 Meaningful human control

Autonomy questions in AI ethics are often framed in terms of "human oversight", but oversight is meaningful only if it is real, competent, and action-guiding. A human who cannot understand the system's role, cannot challenge its output, or is structurally pressured to defer to automation does not exercise meaningful control. Recent scholarship therefore asks whether human oversight is still possible in practice as systems become more complex, adaptive, and deeply embedded in institutional workflows (Holzinger, Zatloukal, and Müller 2025).

Scott Robbins's analysis is particularly helpful for ethics review because it makes clear that "meaningful human control" is not a single design feature but a family of questions: which humans count as relevant decision-makers, what exactly must be controlled, whether control is substantively meaningful rather than merely formal, and whether control extends beyond outputs to training data, inputs, institutional settings, and avenues of contestation. For RECs, that broader framing is valuable because it discourages the superficial equation of human control with the mere presence of a human operator somewhere in the loop (Robbins 2024). Meaningful human control should also be connected to the allocation of responsibility. In AI research, harms may result from distributed decisions across several actors, including model developers, data providers, platform vendors, deployers, research teams, and institutional users. A protocol should therefore identify who is responsible for anticipating, preventing, monitoring, explaining, and remedying reasonably foreseeable harms. For REC review, a "human in the loop" arrangement is insufficient if no responsible actor has the competence, authority, and resources to intervene when the system fails or produces harmful effects.

For RECs, meaningful human control should be analyzed in relation to the concrete task. Does the AI merely support judgment, or does it effectively structure the decision space? Are users given time, training, and authority to disagree with outputs? Is override possible and documented? Are there incentives that make deviation from AI recommendations costly? These questions matter because nominal "human in the loop" arrangements can conceal real deference to automation.

Meaningful control also has a participant-facing dimension. Research participants and affected persons should not be placed in interactions where they must accept AI outputs or AI-mediated decisions without avenues for explanation, challenge, or human recourse. This is especially important in high-stakes or asymmetrical settings such as healthcare, education, employment, law enforcement, and public administration.

4.2 Privacy and data governance

Autonomy is closely connected to informational self-determination. AI systems often rely on large, heterogeneous, and repurposed datasets, which can increase risks of excessive collection, secondary use, re-identification, context collapse, or unjustified inference. Even when data are lawfully processed, researchers should ask whether persons retain meaningful control over how information about them is mobilized, combined, and interpreted. Privacy in AI research is therefore not only about confidentiality. It also concerns the ethical limits of inference: what may legitimately be

inferred about participants or groups from the available data, whether such inferences remain within the purposes participants could reasonably expect, and whether AI-generated conclusions exceed what the data can responsibly support.

RECs should pay particular attention to context. A dataset collected in one relational setting may become ethically problematic when reused in another, especially when sensitive, behavioural, biometric, or health-related data are involved. Generative systems introduce additional concerns because prompts, uploads, and outputs may leak confidential information or intellectual property if researchers use untrusted tools. The European Commission's living guidelines on generative AI in research therefore stress privacy, confidentiality, and intellectual property as central dimensions of responsible use (European Commission 2025a).

Data governance review should cover necessity, minimization, provenance, access control, retention, de-identification limits, vendor dependence, cross-border transfer, and the governance of downstream reuse. It should also ask whether the protocol's promises to participants are compatible with the actual technical and contractual conditions of the AI infrastructure.

4.3 Manipulation, dependency, and behavioural steering

AI systems can shape autonomy not only by replacing judgment but by influencing behaviour. Personalization, nudging, conversational design, recommender logic, affective interfaces, and adaptive feedback can steer users' choices, attention, or trust.

In research contexts, this may affect recruitment, consent, adherence, participant behaviour during the study, or the behaviour of researchers themselves. The ethical issue is not simply whether influence occurs, because all interfaces influence behaviour to some extent, but whether the influence is transparent, proportionate, and compatible with voluntary and reflective agency. AI systems should influence people only in ways they can understand, reasonably resist, and evaluate for themselves. For example: RECs should assess whether the AI system's influence on users or affected persons is transparent, proportionate to the legitimate aim pursued, and compatible with individuals' ability to make informed, voluntary, and considered decisions.

This concern extends to dependency. Researchers and institutions may become dependent on AI systems that are hard to audit, difficult to replace, or embedded in routine workflows. Not every form of dependence is ethically problematic; some reliance on technical infrastructure is unavoidable in contemporary research. Dependence becomes ethically or institutionally concerning when it makes independent verification, replacement, refusal, or correction unrealistic; when commercial or technical infrastructures constrain researchers' methodological choices; when switching costs make ethically preferable alternatives practically unavailable; or when institutional expectations pressure researchers to rely on AI despite unresolved risks. RECs should therefore ask not only whether lock-in exists, but whether its degree is proportionate to the study's aims and accompanied by credible audit, fallback, and exit options.

4.4 Democracy, equity, and power asymmetries

Autonomy in AI research is not only an individual relation between a user and a system; it also depends on the institutional and infrastructural conditions under which that relation is organized. The primary concern remains the agency of participants, users, and affected persons: whether they

can understand, refuse, contest, or seek human review of AI-mediated processes. However, this agency can be constrained by decisions made by researchers and institutions, especially where research relies on commercial or proprietary AI infrastructures controlled by a small number of platform providers, model vendors, or highly specialized institutions. Such infrastructures may determine what data are processed, what explanations are available, what contestation routes exist, and how easily a system can be replaced. RECs should therefore examine whether commercial dependencies and infrastructural asymmetries undermine the practical autonomy of users, participants, or affected communities.

RECs should consider whether the study reproduces problematic asymmetries between developers and users, institutions and participants, or powerful and vulnerable communities. Equity concerns include language exclusion, disability access, unequal digital and AI literacy, and the differential ability to refuse participation or demand human review. A study can therefore raise autonomy concerns even when consent is formally obtained. The deeper question is whether affected persons have meaningful voice and whether the institutional context permits contestation.

4.5 Recommendations

- Require protocols to explain what "human oversight" means operationally: who reviews outputs, with what authority, at what stage, and under what time and workload conditions.
- Ask whether users can meaningfully override or contest AI outputs and whether deviations are institutionally supported rather than discouraged.
- Require a clear data governance plan addressing provenance, necessity, minimization, access control, confidentiality, retention, and vendor dependence.
- Require additional safeguards where AI is used in recruitment, consent support, conversational interaction, or other settings with heightened risks of behavioural steering or manipulation.
- Ask how the protocol addresses digital inequality, accessibility, language exclusion, and differential ability to exercise refusal or contestation rights.
- Where the system may meaningfully shape choices or judgments, ask whether the research design preserves reflective agency for participants, professionals, and researchers.
- Require protocols to assign responsibility for AI-related oversight, intervention, documentation, and remedy. Human oversight should not be accepted as meaningful unless responsible actors have the competence, authority, and resources to act on that responsibility.

5. Research on AI: Safety, Security and Well-Being

5.1 Safety and robustness

Safety concerns whether an AI system behaves acceptably under reasonably foreseeable conditions. Robustness concerns whether performance degrades gracefully when inputs, environments, or user behaviour change. In ethics review, safety is not exhausted by average performance in a development dataset. RECs should distinguish between adversarial conditions during training or development and adversarial conditions in deployment. The first concerns whether the model has been tested against challenging, biased, manipulated, rare, or edge-case data in order to improve robustness. The second concerns whether the system can withstand real-world attempts to game,

manipulate, misuse, or strategically exploit it once it is in operation. Researchers are usually better able to anticipate adversarial conditions during development; adversarial conditions in deployment are often more context-dependent and harder to predict. Nevertheless, researchers should consider reasonably foreseeable forms of misuse, manipulation, and strategic behaviour in light of the intended use context. The practical importance of these issues is clear from recent healthcare reviews, which stress the need for evaluation frameworks that connect validation, usefulness, transparency, and equity to ongoing monitoring after implementation (Wells et al. 2025; De Micco et al. 2025).

Safety should be judged in relation to foreseeable harm. In some research settings, hallucinations, unstable classifications, or silent failure modes may be minor annoyances. In others, the same technical behaviours may create serious welfare, health, reputational, or legal harms. A REC should therefore ask not simply whether the model is accurate, but what happens when it is wrong, how often those errors are likely to occur in practice, who notices them, and what mitigation pathway exists.

The burden of proof should rise with the stakes of the use case. Studies involving participant-facing decision support, health, behavioural intervention, social scoring, biometric inference, or vulnerable populations should be expected to present more substantial evidence of validation, fallback procedures, escalation pathways, and post-approval monitoring.

5.2 Security

AI security includes risks such as model theft, prompt injection, data poisoning, adversarial attacks, membership inference, extraction of sensitive information, and unauthorized access to model inputs or outputs. These are ethical concerns because security failures can harm participants, compromise confidentiality, distort research findings, and undermine trust. In some cases, weak security also converts otherwise proportionate research into unethical research because the residual risk becomes unjustifiable. Security review should be proportionate but concrete. Researchers should describe threat models, access controls, third-party dependencies, incident response, logging, patching, and the division of responsibility between the research team and external vendors. Where systems process sensitive data or are used in consequential settings, ethics review should expect clearer evidence that security has been considered from the design stage onward. Recent reviews emphasize that secure and trustworthy AI in healthcare depends not only on technical protection, but also on governance structures that connect innovation to accountability and ethical foresight (Khan et al. 2025).

5.3 Explainability and the need to understand AI

Explainability is sometimes discussed as if it were a universally required property. In fact, what matters ethically is whether the available understanding is sufficient for the specific review and use context. RECs should ask why understanding is needed in the case at hand: to contest decisions, to validate outputs, to detect failure, to assign responsibility, to support informed interaction, or to ensure that the model aligns with domain knowledge. A simple but opaque model may sometimes be more ethically acceptable than a badly explained but deeply consequential system. Conversely, a

highly accurate system may remain ethically problematic if its operation cannot be interrogated where explanation is needed for accountability or safety.

The need to understand AI is therefore linked to the role the system plays in the research design. If the output merely suggests low-stakes hypotheses that are later independently verified, deep mechanistic explanation may be less important. If the output meaningfully structures participant treatment, recruitment, moderation, or resource allocation, then interpretability and traceability may become central ethical requirements. Here, AIOLIA's work on operationalizing transparency and accountability is useful because it connects explanation not to abstract intelligibility alone but to practical traceability, communication, and responsibility (Bayerl et al. 2026).

5.4 Dual use and misuse

AI research often has dual-use potential. A model, dataset, or method developed for beneficial purposes may later be repurposed for surveillance, deception, discrimination, cyber offence, or other harmful applications. Generative systems add further concerns because they can facilitate scale, speed, and plausibility in the creation of synthetic text, code, images, voice, or video. Dual-use risk does not automatically render research unethical, but it does require anticipatory reflection and proportionate safeguards. RECs should ask whether the protocol includes foreseeable misuse scenarios, whether access to models or data is appropriately governed, and whether publication or release decisions have been ethically considered. In some cases, staged access, redaction of sensitive implementation detail, or stronger governance around reuse may be appropriate. The goal is not to suppress research by default, but to ensure that the pathway from beneficial use to harmful reuse has been taken seriously.

5.5 Beneficence and non-maleficence

Classical research ethics principles remain relevant in AI research, but they require translation. Beneficence demands more than speculative claims of innovation or efficiency. Researchers should demonstrate plausible pathways by which the study could generate benefits, for whom those benefits are likely to accrue, and how those benefits compare with foreseeable risks. Non-maleficence requires more than avoiding obvious physical injury; it includes avoiding informational harm, dignitary harm, discriminatory harm, dependency, institutional distortion, and erosion of trust.

Recent practical frameworks, including FAIR-AI and European guidance on AI in science, support a view of beneficence that is inseparable from validation, governance, and post-implementation review. A system that appears promising but cannot be safely integrated, monitored, or contested may fail the ethical test even if its performance metrics are impressive (Wells et al. 2025; European Commission 2025a).

5.6 Recommendations

- Require validation evidence proportionate to the stakes of the use case, including plans for monitoring system drift, failure, and harmful unintended consequences after initial approval where relevant.

- Ask what happens when the system is wrong, who detects the error, and what escalation or fallback process is available.
- Require a security section covering threat models, access control, third-party dependencies, incident response, and protection of sensitive data and outputs.
- Where explainability is invoked, require researchers to specify for whom explanation is needed and what kind of understanding is sufficient for the use case.
- Require protocols to address foreseeable misuse or dual-use pathways, especially when models, datasets, or outputs may be repurposed in harmful ways.
- Do not accept vague claims of social benefit as sufficient. Ask who benefits, under what assumptions, and with what distribution of burdens.

6. Research with AI: Research Integrity, AI, and the Shapes of Contemporary Science

6.1 Research integrity and AI

The use of AI in research practice raises questions that are partly continuous with older integrity concerns and partly new. Long-standing duties of reliability, honesty, respect, and accountability remain in force, but generative AI and other AI tools make it easier to outsource drafting, coding, summarizing, classification, and review tasks in ways that blur the location of judgment and responsibility. The European Code of Conduct for Research Integrity remains a foundational reference, and the European Commission's living guidelines on generative AI in research build directly on it by articulating responsibilities for researchers, research organizations, and funders (ALLEA 2023; European Commission 2025a).

From an integrity perspective, the core question is not whether AI was used at all, but whether its use was transparent, verifiable, and consistent with responsible scholarship. A researcher may appropriately use AI for language support or coding assistance while remaining fully responsible for checking outputs, disclosing substantial use where relevant, protecting confidential material, and preserving the integrity of the research process. Problems arise when AI use obscures provenance, introduces fabricated or inaccurate material, compromises confidential inputs, or erodes the researcher's own control over claims and methods (Vasconcelos and Marušić 2025; Knöchel et al. 2025; Helmy et al. 2025).

A stronger research-ethics literature now gives these concerns sharper contours. Resnik and Hosseini argue in AI and Ethics that the established norms of science remain in force under AI use, but that they require more explicit guidance on bias control, disclosure, stakeholder engagement, synthetic data, authorship, and education. In Research Ethics, Hosseini, Resnik, and Holmes similarly reject blanket bans on large language models in scholarly writing and defend a transparency-oriented disclosure model tied to accountability and fair allocation of credit. Resnik and Hosseini later refine this position in Accountability in Research by distinguishing cases in which disclosure should be mandatory, optional, or unnecessary, a distinction that is highly relevant for REC expectations

concerning protocol drafting, analytic assistance, and publication planning (Resnik and Hosseini 2025; Hosseini, Resnik, and Holmes 2023; Resnik and Hosseini 2026).

6.2 Research integrity and epistemic principles

Research integrity is not only procedural. It is also epistemic. The use of AI can affect how evidence is generated, filtered, interpreted, and justified. If researchers rely heavily on systems that produce plausible but weakly grounded text, code, or synthetic data, they may degrade the evidential quality of their work even while increasing productivity. Similarly, if AI systems shape the framing of questions, search strategies, coding categories, or hypothesis spaces, then epistemic bias can enter the research process before any explicit result is generated.

Recent work by Resnik, Hosseini, and Hauswald extends this concern by asking how scientific values should be protected as AI systems become more autonomous within research workflows. The more such systems participate in hypothesis generation, evidence ranking, or intermediate evaluative steps, the more important it becomes to preserve identifiable human responsibility, contestability, and alignment with the human values that justify scientific research in the first place (Resnik, Hosseini, and Hauswald 2026).

The emergence of AI agents adds a further layer to these concerns. AI agents used in research, including social and biomedical research, may not only generate outputs but also plan tasks, select tools, interact with datasets, trigger intermediate steps, and make preliminary research decisions. This can support scientific values such as efficiency, discovery, and methodological exploration. At the same time, it may threaten transparency, epistemic accountability, validity, and respect for participants if human oversight is weak or merely formal. Researchers should therefore remain responsible for defining the aims, constraints, data uses, and acceptable forms of automation rather than treating agentic AI as an independent research actor. Particular attention should be paid to alignment: AI agents used in research should be aligned not only with task-specific goals, but also with scientific standards, ethical norms, participant protections, and institutional accountability.

The relevant concern is not only fraud or fabrication. It is also the more ordinary erosion of methodological vigilance when AI outputs are accepted too quickly because they look competent or save time. The literature increasingly warns that generative AI may challenge established concepts of originality, justification, and scholarly responsibility unless institutions insist on human verification and context-sensitive disclosure (Vasconcelos and Marušić 2025; European Commission 2025a).

6.3 Authorship, responsibility, and accountability

A central integrity issue concerns authorship and responsibility. Current journal and editorial guidance is clear on one point: AI systems cannot be authors because they cannot take responsibility for the work, approve the final version, or be accountable for integrity. This is reflected in the current recommendations of the International Committee of Medical Journal Editors and in COPE guidance. Human authors remain fully responsible for the content they submit, including material generated or supported by AI (ICMJE 2025; COPE 2023).

For RECs, the authorship issue is broader than publication policy. It concerns responsibility throughout the research lifecycle. If AI is used to generate protocol language, code, analytic suggestions, participant-facing content, or review summaries, the research team should specify who checks those outputs, how they are checked, and who remains accountable for them. The European Commission's guidance additionally advises against the use of generative AI for assessment or evaluation of scientific content where human judgment is essential, a point of direct relevance to peer review, grant evaluation, and some forms of REC support (European Commission 2025a).

The practical standard should be that AI may support work, but cannot bear responsibility for it. Accountability must remain human, named, and reviewable.

6.4 Deskilling and upskilling

AI may both support and erode expertise. On the one hand, tools can help researchers work across languages, automate repetitive tasks, and gain access to analytical capacities that would otherwise be unavailable. On the other hand, habitual reliance on AI may weaken domain judgment, writing ability, coding competence, interpretive skill, or methodological sensitivity. This concern is especially acute for early-career researchers and for institutional environments that normalize speed and output over reflective understanding.

Deskilling is ethically relevant because research quality depends not only on outputs but on trained human capacities. If AI use gradually substitutes for understanding, then the apparent efficiency gain may come at the cost of weaker scholarship and poorer supervision. AIOLIA explicitly identifies over-reliance and deskilling as practical governance concerns, and current research-integrity debates similarly emphasize the need for training and deliberate human competence development rather than uncritical tool adoption (Bayerl et al. 2026; Knöchel et al. 2025).

At the same time, responsible upskilling is possible. Institutions can train researchers to use AI critically, to disclose use proportionately, to validate outputs, to protect confidential materials, and to recognize when AI should not be used. RECs may therefore consider not only whether a protocol uses AI, but whether the research team is demonstrably competent to use it responsibly. In short, AI can sharpen research practice or quietly hollow it out. The key ethical question is whether researchers use AI as a scaffold for developing and extending expertise, or as a substitute that erodes the skills needed to judge, challenge, and understand its outputs. RECs should therefore look for training, verification routines, and task boundaries that preserve the human competences on which research quality depends.

6.5 Missing understanding and black boxes in science

A final concern is that AI can produce useful outputs without yielding meaningful scientific understanding. In some settings, black-box systems may improve prediction while leaving underlying mechanisms opaque. This may be acceptable for certain exploratory or assistive purposes, but problematic when the research aim includes explanation, causal understanding, normative justification, or clinically or socially accountable decision-making. The ethical issue is not that all science must be mechanistically transparent, but that the role of understanding in the study should be explicit and not silently displaced by predictive convenience.

RECs should therefore ask whether the use of AI is compatible with the epistemic aims claimed in the protocol. If a project claims to improve understanding, explanation, or theory-building, yet relies on tools whose contribution cannot be meaningfully interrogated, there may be a mismatch between method and stated aim. Such mismatches can become integrity issues if they are hidden behind inflated claims of innovation or objectivity.

6.6 Recommendations

- Require transparent disclosure of substantial AI use in the research process, especially where it affects drafting, coding, analysis, synthesis, image generation, or evaluation.
- Require researchers to specify who checks AI-generated or AI-supported outputs and by what process.
- Treat human accountability as non-delegable: AI systems cannot be authors and cannot bear responsibility for the integrity of the work.
- Require safeguards for confidential, personal, proprietary, or unpublished material before generative AI tools are used in the research workflow.
- Ask whether AI use is compatible with the epistemic aims of the project, including explanation, interpretation, and reproducibility.
- Where reliance on AI may contribute to deskilling, require evidence of training, supervision, or workflow design that preserves human competence and reflective judgment.

7. Pulling the ethics review together

The preceding sections are intended to help RECs and researchers structure their ethical assessment rather than to produce an exhaustive checklist. In practice, an acceptable review integrates several questions at once: what kind of AI is involved, what role it plays in the study, who may be harmed, how control is distributed, whether the system is safe and secure enough for the use case, and whether the conduct of research remains transparent, accountable, and epistemically reliable. This integrated approach reflects both the iRECS recommendation to move beyond one-time compliance and the AIOLIA emphasis on operationalizing principles in context-specific ways (Aucouturier and Grinbaum et al. 2023; Bayerl et al. 2026).

7.1 Key questions for RECs

Dimension	Core REC question	Evidence to request
System and context	What kind of AI is used, for what task, and with what stakes?	System description; intended use; development stage; deployment context
Fairness	Who may be disadvantaged, and how was bias assessed?	Data provenance; subgroup evaluation; fairness rationale

Autonomy and control	Who is able to question or override the system?	Human oversight plan; recourse and contestation pathways
Safety and security	What happens when the system fails or is attacked?	Validation; monitoring; fallback procedures; threat model
Data governance	Are the data and infrastructure ethically and legally fit for purpose?	Data management plan; vendor and access information
Research integrity	How is AI used in the conduct of research itself?	Disclosure plan; verification workflow; authorship/accountability statement

7.2 What researchers should include in the protocol

A protocol involving AI should provide enough information for the committee to judge the ethical acceptability of the system in context. At a minimum, researchers should describe the AI system and its intended role; the data sources and their provenance; the populations represented and potentially affected; the performance and validation strategy; the fairness rationale; the plan for human oversight; the security and incident-response approach; and the conditions of post-approval monitoring where relevant. If generative AI is used in the conduct of the research, the protocol should also explain what was used, for what purpose, what safeguards were in place, and how human verification occurred.

It is often useful for RECs to require a short annex specifically devoted to AI-related ethical and governance matters. Such an annex can prevent the most ethically relevant information from being dispersed across data management, methodology, and dissemination sections.

7.3 Red flags

- Vague statements that the project uses “AI” without specifying the model type, task, data, or deployment setting.
- Claims that the system is “fair”, “robust”, or “trustworthy” without evidence, metrics, or justification.
- No discussion of affected groups beyond recruited participants, despite plausible wider social impact.

- Nominal references to human oversight without an operational account of who can review, challenge, or override outputs.
- No plan for monitoring drift, incidents, or harmful downstream effects in dynamic or consequential systems.
- Use of generative AI in protocol drafting, coding, analysis, or dissemination without disclosure and verification procedures.
- Dependence on third-party tools or vendors without explanation of confidentiality, security, access control, or intellectual property implications.
- Overstated claims about benefit or innovation with limited evidence of validation or proportionality.

7.4 Cross-cutting recommendations

- Ensure proposals distinguish clearly between research on AI and research with AI, while allowing for mixed cases.
- Ensure that the role of the AI system in the research design is described precisely enough for ethical review.
- Require proportionate evidence on data provenance, fairness, validation, oversight, security, and post-approval monitoring.
- Require a realistic account of human responsibility at every stage of the project, including publication and dissemination.
- Where AI-related expertise is lacking within the REC, seek internal or external expert input rather than defaulting to superficial review.
- Move beyond one-time compliance where the system is dynamic, adaptive, or likely to create ethically relevant effects after initial approval.
- Encourage ethics-by-design and ongoing ethical reflection throughout the project lifecycle, in line with insights from sister projects such as iRECS and AIOLIA.

8. References and further resources

ALLEA. 2023. The European Code of Conduct for Research Integrity. Revised edition. Berlin: ALLEA.

Aucouturier, Etienne, and Alexei Grinbaum. 2024. Reviewing the Ethics of AI in Health and Healthcare. iRECS Policy Brief.

Aucouturier, Etienne, Alexei Grinbaum, et al. 2023. iRECS D2.2: Recommendations to Address Ethical Challenges from Research in New Technologies.

Bagattini, Alexander, and Chen Shao. 2025. “Prejudiced Futures? Algorithmic Bias in Time Series Forecasting and Its Ethical Implications.” arXiv preprint arXiv:2512.01877. https://arxiv.org/abs/2512.01877.

Bashkin, Osnat. 2025. “Patient Perspectives on Artificial Intelligence in Healthcare: A Global Scoping Review of Benefits, Ethical Concerns, and Implementation Strategies.” International Journal of Medical Informatics 203: 106007. https://doi.org/10.1016/j.ijmedinf.2025.106007.

Bayerl, Petra Saskia, Erin Lawlor, Maria Teresa Maris, Denis Miorandi, Gintarė Pekšys, Miroslav Bjelica, Katarina Stojšin, Mariya Anastasova, Oliver Smith, Arturo Henestrosa, Ivan Yamshchikov, M. A. R. Bak, and Babak Akhgar. 2026. Operational Ethics Guidelines on Use Cases Related to Human Behaviour and Cognition. AIOLIA Deliverable D3.1.

Chen, Richard J., Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. 2023. “Algorithmic Fairness in Artificial Intelligence for Medicine and Healthcare.” Nature Biomedical Engineering 7 (6): 719–42. https://doi.org/10.1038/s41551-023-01056-8.

Committee on Publication Ethics (COPE). 2023. “Authorship and AI Tools.” COPE position statement.

De Micco, Francesco, Gianmarco Di Palma, Davide Ferorelli, Anna De Benedictis, Luca Tomassini, Vittoradolfo Tambone, Mariano Cingolani, and Roberto Scendoni. 2025. “Artificial Intelligence in Healthcare: Transforming Patient Safety with Intelligent Systems—A Systematic Review.” Frontiers in Medicine 11: 1522554. https://doi.org/10.3389/fmed.2024.1522554.

European Commission. 2021. Ethics by Design and Ethics of Use Approaches for Artificial Intelligence. Brussels: European Commission.

European Commission. 2025a. Living Guidelines on the Responsible Use of Generative AI in Research. 2nd ed., April 2025. Brussels: Directorate-General for Research and Innovation, ERA Forum Stakeholders’ Document.

European Commission. 2025b. “AI Act: Questions and Answers” and related implementation materials on the EU regulatory framework for artificial intelligence. Brussels: European Commission.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Helmy, Mohamed, Lingling Jin, Amr Alhossary, Tamer Mansour, Diogo Pellagrina, and Kumar Selvarajoo. 2025. “Ten Simple Rules for Optimal and Careful Use of Generative AI in Science.” PLoS Computational Biology 21 (10): e1013588. https://doi.org/10.1371/journal.pcbi.1013588.

High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. Brussels: European Commission.

High-Level Expert Group on Artificial Intelligence. 2020. Assessment List for Trustworthy AI (ALTAI). Brussels: European Commission.

Holzinger, Andreas, Karel Zatloukal, and Heimo Müller. 2025. “Is Human Oversight to AI Systems Still Possible?” New Biotechnology 85: 59–62. https://doi.org/10.1016/j.nbt.2024.12.003.

Hosseini, Mohammad, David B. Resnik, and Kristi Holmes. 2023. "The Ethics of Disclosing the Use of Artificial Intelligence Tools in Writing Scholarly Manuscripts." Research Ethics 19 (4): 449–65. https://doi.org/10.1177/17470161231180449.

International Committee of Medical Journal Editors (ICMJE). 2025. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.

Khan, Muhammad Mohsin, Noman Shah, Nissar Shaikh, Abdulnasser Thabet, Talal Alrabayah, and Sirajeddin Belkhair. 2025. "Towards Secure and Trusted AI in Healthcare: A Systematic Review of Emerging Innovations and Ethical Challenges." International Journal of Medical Informatics 195: 105780. https://doi.org/10.1016/j.ijmedinf.2024.105780.

Knöchel, Tim-Dorian, Konrad J. Schweizer, Oguz A. Acar, Atakan M. Akil, Ali H. Al-Hoorie, Florian Buehler, Mahmoud M. Elsherif, Alice Giannini, Evelien Heyselaar, Mohammad Hosseini, Vinodh Ilangovan, Marton Kovacs, Zhicheng Lin, Meng Liu, Anco Peeters, Don van Ravenzwaaij, Marek A. Vranka, Yuki Yamada, Yu-Fang Yang, and Balazs Aczel. 2025. "Core Principles of Responsible Generative AI Usage in Research." AI and Ethics 5 (6): 6371–77. https://doi.org/10.1007/s43681-025-00768-8.

Koçak, Burak, et al. 2025. "Bias in Artificial Intelligence for Medical Imaging: Fundamentals, Detection, Avoidance, Mitigation, Challenges, Ethics, and Prospects." Diagnostic and Interventional Radiology 31: 75–88.

National Institute of Standards and Technology (NIST). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: NIST.

National Institute of Standards and Technology (NIST). 2024. Generative Artificial Intelligence Profile. Gaithersburg, MD: NIST.

OECD. 2024. OECD Recommendation of the Council on Artificial Intelligence. Paris: Organization for Economic Co-operation and Development.

Resnik, David B., and Mohammad Hosseini. 2025. "The Ethics of Using Artificial Intelligence in Scientific Research: New Guidance Needed for a New Tool." AI and Ethics 5: 1499–1521. https://doi.org/10.1007/s43681-024-00493-8.

Resnik, David B., and Mohammad Hosseini. 2026. "Disclosing Artificial Intelligence Use in Scientific Research and Publication: When Should Disclosure Be Mandatory, Optional, or Unnecessary?" Accountability in Research 33 (2): 2481949. https://doi.org/10.1080/08989621.2025.2481949.

Resnik, David B., Mohammad Hosseini, and Rico Hauswald. 2026. "Autonomous Artificial Intelligence, Scientific Research, and Human Values." AI and Ethics 6 (1): 141. https://doi.org/10.1007/s43681-025-00908-0.

Robbins, Scott. 2024. "The Many Meanings of Meaningful Human Control." AI and Ethics 4: 1377–88. https://doi.org/10.1007/s43681-023-00320-6.

Schmidt, M. W., Bagattini, A., Pfister, J., Martin, D., & Hillerbrand, R. (2026: forthcoming). KI-Ethik und KI-Verordnung. In J. Wendt & D. Wendt (Hrsg.), Verordnung über künstliche Intelligenz: KI-VO. Baden-Baden: Nomos.

Teo, Sue Anne, Nicoleta Kyosovska, and Abril Armengol. 2025. AIOLIA D2.2: Report on the Selection of Ethical Principles and Values. June 2025.

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.

Vasconcelos, Sonia, and Ana Marušić. 2025. "Gen AI and Research Integrity: Where to Now?" EMBO Reports 26 (8): 1923–28. https://doi.org/10.1038/s44319-025-00424-6.

Wells, Brian J., Hieu M. Nguyen, Andrew McWilliams, Matt Pallini, Amy Bovi, Andrew Kuzma, Justin Kramer, Shih-Hsiung Chou, Timothy Hetherington, Patricia Corn, Yhenneko J. Taylor, Audrey Cuison, Mary Gagen, McKenzie Isreal, and FAIR-AI Consortium. 2025. "A Practical Framework for Appropriate Implementation and Review of Artificial Intelligence (FAIR-AI) in Healthcare." npj Digital Medicine 8: 514. https://doi.org/10.1038/s41746-025-01900-y.

www.changer-project.eu

This project has received funding from the European Commission Horizon WIDERA under Grant Agreement No 101131683