SECTION «PHILOSOPHICAL SCIENCES»

MORAL OUTSOURCING: GENERATIVE ARTIFICIAL INTELLIGENCE AND THE SUSTAINABILITY OF VALUE TRANSMISSION

Olesia Pankiv[1]

DOI: https://doi.org/10.30525/978-9934-26-715-4-44

Abstract. The transfer of moral judgment to generative language models is a new cultural practice whose consequences reach beyond the ethics of technology. When a person asks a system not for the facts relevant to a choice but for the choice itself, the formal authorship of the decision remains with the person while part of the work that constitutes judgment is performed elsewhere. This study designates such a situation as moral outsourcing and examines it as a cultural process bearing on the intergenerational transmission of values. *The purpose* of the paper is to conceptualise moral outsourcing as a factor affecting the sustainability of value transmission, to outline the risks associated with it, and to identify the conditions under which interaction with a generative model supports rather than weakens the moral competence of the subject. *Methodology* of the study is interdisciplinary. Conceptual analysis establishes the boundaries of the notion and distinguishes it from moral consultation, moral enhancement and moral automation. Hermeneutic reconstruction is applied to the philosophical tradition of autonomy and judgment, from the Kantian distinction between autonomy and heteronomy to the postphenomenology of technological mediation. A critical review of empirical research published over the past three years establishes the current state of knowledge, while cultural analysis situates the phenomenon within the problematics of value transmission. The study is theoretical and conceptual in character: the empirical works it draws upon are treated as material for critical examination rather. *Results* of the study show that the

[1] Candidate of Philosophical Sciences, Associate Professor,
Mykola Lysenko Lviv National Music Academy, Ukraine

© Olesia Pankiv

principal risk lies not in the incorrectness of an individual recommendation but in three structural effects. The first is moral deskilling: moral competence is partly constituted by practical skills acquired through repeated exercise, and technologies that take over morally loaded actions reduce the occasions for it. The second is the asymmetry of cultural representation: the structure of data available to a model may privilege certain languages and modes of moral argumentation, so that together with an answer the user may also adopt a way of seeing the situation. The third is the weakening of the intergenerational channel, as moral guidance is sought from an instance that shares neither the community's lived history nor its collective memory, and bears no part in its future. The risks are ordered into a four-level typology, and none of them is inevitable, since each is realised only under particular conditions of interaction. *Practical implications.* Interaction with a generative model can sustain rather than erode moral competence provided that the subject formulates a position before consulting the system; that the request is for counter-arguments rather than for a conclusion; that several systems are consulted, since their low mutual agreement reveals that what is offered is a position and not a truth; that educational formats are built on exercise rather than on information about risks; and that developers disclose the post-training and alignment stages, in which a substantial part of the systems' moral biases appears to originate. *Value/originality.* The paper shifts the question from the moral correctness of particular answers given by artificial intelligence to the cultural consequences of the act of delegation itself, and proposes the sustainability of value transmission as an analytical category. The Ukrainian wartime context is introduced not as an illustration but as a particularly revealing case, in which moral dilemmas of high social stake are combined with their uneven representation in global language data.

INTRODUCTION

A few years ago a request addressed to a machine along the lines of “Did I act rightly?” might have seemed an episode from a science-fiction film. Today it is an ordinary practice. A person opens a chat, describes a conflict with someone close, a difficult choice or a doubt about their own conduct, and asks not merely for help in making sense of the circumstances but for an answer to a far more demanding question: what is the right thing

to do? Generative artificial intelligence is gradually entering the sphere of human life traditionally regarded as the most personal and the least transferable – the sphere of moral judgment.

There is something of fundamental importance in this transition. For centuries people sought moral advice from those who belonged to their world: from parents and elders, friends and teachers, priests, literary characters and, finally, their own conscience. Such advice was never merely the delivery of a ready answer. It involved conversation, disagreement, the possibility of censure, the need to explain one's position and to withstand another's gaze. A moral decision took shape not only through the advice received but in the very process of seeking it. Today part of this work is increasingly taken over by a system that is available around the clock, demands no reciprocity, passes no judgment and almost always has a ready, well-formulated answer.

Hence a question arises that extends far beyond the technical capacities of artificial intelligence. What happens to a person who transfers to a machine not merely the execution of a task but part of the work of their own judgment? And, more difficult still, what happens to a community when such a practice becomes widespread? Moral judgment is not a purely individual act. It determines what counts as just and unjust, worthy and unworthy; through it are formed the ways of speaking about responsibility, guilt, obligation and forgiveness. Moral language does not exist outside culture: it is passed from generation to generation together with experience, memory and the practices of shared life.

It is essential that the problem here does not reduce to the quality of the advice. Even a correct, persuasive and morally acceptable answer may have a paradoxical consequence – it may relieve the person of the need to travel the path to that decision themselves. In debates on the ethics of artificial intelligence it is natural to ask whether algorithms are capable of morally correct decisions and what biases the training data contain. Beyond that approach, however, another and no less important question remains: does the very practice of turning to an algorithm change our capacity to judge, to doubt, to argue and to pass our own moral orientations to the generations that follow? What is at issue is the boundary between assistance and substitution, between argued dialogue and a ready answer. If this boundary shifts imperceptibly, the system influences not so much what a

person considers right as the way in which they become accustomed to determining what is right.

In the Ukrainian context the question becomes especially pressing. The moral experience of Ukrainian society in recent years has been shaped to a considerable degree by the war and by situations that do not fit the familiar patterns of everyday choice. Questions about the limits of personal responsibility, obligation to the community, forced choice, forgiveness, memory and the relation between personal safety and collective duty are not abstract intellectual exercises: they arise in real life, carry a high cost and often admit of no unconditionally right solution. The moral vocabulary required to make sense of such experience is formed within the community – in families, volunteer circles, public discussion, art and everyday conversation. A generative system may be extremely persuasive, yet it does not share the community's lived history, its collective memory, or responsibility for the consequences of its advice.

UNESCO's Recommendation on the Ethics of Artificial Intelligence stresses the need to protect cultural diversity under conditions in which automated systems are becoming widespread [23, p. 24], and in Ukrainian normative discourse the ethical dimension of the development of these technologies likewise occupies a place of principle [16]. Yet the cultural problematics of artificial intelligence is not exhausted by questions of representation or linguistic diversity. Culture is reproduced not only through the preservation of texts, norms and symbols but also through the capacity of each new generation to rethink contradictory experience anew, to pose moral questions and to assume responsibility for its own choices. If this practice changes, the very mechanism of cultural transmission may change with it. For this reason moral outsourcing is treated here as a cultural process and not solely as a problem in the ethics of technology.

At the same time it is important to distinguish what has already been confirmed by empirical research from what remains a well-grounded hypothesis. The available data give grounds for speaking of the influence of language models on moral judgment in the short term, whereas questions of a long-term departure from previously acquired moral attitudes, of changes in cultural frames and of the transformation of intergenerational value transmission require further verification. The significance of the problem is determined not by fear of a new technology but by the need to

understand its quiet, everyday and cumulative effects. The greatest cultural transformation may occur not when a machine begins to make decisions instead of a person, but when a person gradually ceases to notice that certain decisions are no longer of their own making.

1. The Current State of Research on the Problem

The current state of research warrants separate consideration, since a substantial field of interdisciplinary research has already formed around human interaction with generative artificial intelligence. At the centre of international scholarly attention are the biases of models conditioned by the structure of the data used for training [4, p. 613-617], normative constraints on the use of such systems [23, p. 24], and the psychological mechanisms of trust in algorithmic recommendations – automation bias [28, p. 998-1003], algorithm appreciation [21, p. 92-99] and the opposing algorithm aversion [7, p. 118-123]. A separate strand of research is found in the philosophy of technology, which has argued against regarding a technical artefact as a neutral mediator [36, p. 127-133; 33, p. 41-57].

A considerable part of this discourse unfolds around the system itself: how correct its answers are and what biases the training data contain. In the moral sphere the unevenness of attention is especially noticeable. Research attests that the advice of language models can shift users' judgments, while users themselves underestimate the degree of this influence [17]. At the same time, the answers of language models are rated no lower than those of professional ethicists [8], and the influence of a model on users' judgments may be comparable to that of a human being, despite a lower declared level of trust in it [19]. In parallel, the properties of the systems themselves are studied: their systematic biases [5] and the divergence of judgments across models [24]. The object of analysis here is consistently the content of the advice rather than the practice of requesting it.

A line of research that raises the question of the subject, and not only of the system, is also present in international discourse, although it is considerably less developed. Its classic beginning is J. Weizenbaum's warning about decisions that ought not to be transferred to computing systems – not because they would fail to cope, but because the transfer itself is a loss [35, p. 188-201]. This line is continued most consistently by S. Vallor in the notion of moral deskilling [30, p. 110-118], and, on a somewhat different

plane, by S. Turkle [29, p. 2135]. The opposing normative position is defended by A. Giubilini and J. Savulescu, for whom an external system free of emotional bias brings the subject closer to the standpoint of an ideal observer [9, p. 174-182]. Yet none of these lines carries the problem to the level of cultural transmission. A fourth strand concerns the cultural variability of morality itself. J. Henrich and co-authors have argued that the populations from which most behavioural research draws its participants sit at one end of the human range rather than at its centre [12, p. 63-72], and the Moral Machine experiment has recorded stable regional clusters of moral preference [2, p. 61-63]. So far, however, these findings have not been brought into contact with the question of moral outsourcing: cultural variation and the influence of systems on moral evaluation are studied apart from one another.

Ukrainian research already possesses work that provides a theoretical foundation for making sense of digital transformations. A general framework for the philosophical interpretation of digitalisation is presented in the works of V. Voronkova and V. Nikitenko, which develop the conception of the digital human being and digital society and analyse the accompanying shifts in values [34, p. 63-92]. From the perspective of cultural studies, the influence of digital technologies on cultural communication is examined by I. Shevel [26], questions of digital inclusion are studied by I. Shvets [27], and the ethical interpretation of machine learning technologies is addressed in [22]. A significant body of Ukrainian research is concentrated, however, on normative and legal questions and on academic integrity, and does not cover situations in which a user turns to a system in order to determine their own position in a situation of moral choice.

Closest to the subject of the present study is the conception of cognitive outsourcing proposed by O. Kryvonos and M. Kozlovets. The authors draw attention to the illusion of competence, in which the subject perceives a result generated by a system as their own intellectual achievement without having passed through the stage of cognitive effort, which may contribute to a weakening of the capacity to produce decisions independently [18, p. 50]. Equally important is their distinction between technological assistance and the substitution of thinking: the boundary between these two modes of using technology runs through the preservation of the subject's intellectual autonomy [18, p. 51].

Carrying this analytical framework across from intellectual activity to moral choice, however, requires a fundamental qualification. A cognitive task can in many cases be assessed by criteria of correctness or error. Moral judgment is of a different nature: it arises in a situation of conflicting values, presupposes no universal algorithm of right decision, and is inseparably bound up with the subject's responsibility for their own choice. In the moral sphere, therefore, the question is not so much whether the subject has received the right answer as whether they remain the author of the judgment that underlies their action.

This distinction makes it possible to define the research gap more precisely. The available works furnish instruments for analysing the digital human being, algorithmic non-neutrality, cognitive outsourcing and the ethical risks of artificial intelligence, yet they do not constitute an integral model for analysing a situation in which a generative system becomes a mediator in the production of moral judgment. Still less studied is the next level of the problem – the possible influence of such a practice on cultural transmission, that is, on the capacity of a community not only to preserve values but also to reproduce the mechanisms by which they are interpreted and passed on. The Ukrainian context is in this respect not peripheral but a particularly promising space of inquiry: here existing work on digital transformation coincides with an intensive rethinking of values and moral orientations, which makes the question of how they are reproduced especially sensitive.

The purpose of the study is to conceptualise moral outsourcing as a factor affecting the sustainability of value transmission, to outline the risks associated with it, and to identify the conditions under which interaction with a generative model supports rather than weakens the moral competence of the subject.

Methodologically the study is interdisciplinary and combines four procedures. Conceptual analysis makes it possible to define the conceptual boundaries of moral outsourcing. Hermeneutic reconstruction is employed in working with the philosophical tradition of autonomy and judgment. A critical review of empirical research conducted between 2023 and 2026 in psychology, cognitive science and the ethics of computing establishes the state of knowledge on the question. Cultural analysis situates the phenomenon within the problematics of value transmission and cultural

sustainability. It is also important to define the status of the study itself: it is theoretical and conceptual in character, and the empirical works on which the analysis rests are treated as material for critical examination.

2. Outsourcing of Judgment: Conceptual Boundaries

The term “moral outsourcing” designates a situation in which a subject faced with the necessity of moral choice turns to an external automated instance not for the information required for that choice, but for a ready variant of the decision or for its justification. It should be noted that the term is used in the literature to describe different configurations of responsibility. R. Chowdhury, in particular, employs it to denote the practice of shifting responsibility for the consequences of technological decisions from people and organisations onto the technological systems themselves [6]. In more recent literature, however, the term is also applied directly to the delegation of an ethical decision to artificial intelligence. It is in this sense that M. Z. Hassen uses it, linking such a practice with the weakening of human moral reasoning and the erosion of accountability [11]. In the present study moral outsourcing is used in the second of these senses.

Terminological demarcation is necessary because outsourcing is not synonymous with delegation in general. Delegation is a broader notion and may denote the transfer of an action to a technical artefact without any presumption that the subject loses the capacity to judge [20, p. 232-241]. Moral outsourcing, by contrast, concerns a situation in which an external means takes over part of the work that constitutes judgment itself. Alongside it, it is useful to consider the notion of moral deskilling proposed by S. Vallor to describe the weakening of moral skills under the influence of technologies that take over the performance of morally significant actions [30, p. 110-118; 31, p. 155-174]. Value transmission here denotes the process of passing on and reproducing values within a culture; its intergenerational dimension draws on the conception of transmission belts developed by U. Schönpflug [25, p. 176-182]. The notion of the sustainability of value transmission, in turn, is formed by analogy with the intergenerational logic of sustainable development [37, p. 43], and refers to a community's capacity to reproduce the conditions under which the transfer of values to subsequent generations is not interrupted.

Moral outsourcing should be distinguished from at least three adjacent phenomena. The first is moral consultation, in which the subject seeks another person's position but remains within a space of shared moral interaction: the adviser may object, seek clarification, demand justification and, ultimately, be held responsible for the advice given. The second is moral enhancement, that is, the use of external means to develop one's own capacity to judge; here technological assistance aims at strengthening competence rather than replacing it. The third is moral automation, in which the system in effect makes the decision without human participation in the particular act of choice. Outsourcing occupies an intermediate position between these poles: the person formally remains the subject of the decision while transferring part of the judgment itself to an external system.

It is this intermediate position that determines the specificity of the phenomenon. Formally the subject retains the marks of authorship: they formulate the request, receive an answer and may accept or reject it. In fact, however, what constitutes the inner work of moral judgment is placed outside their immediate activity: the weighing of grounds, the correlation of conflicting obligations, the selection of language for describing a conflict of values. Analytically it is useful to distinguish here between the outsourcing of a particular decision and the outsourcing of a moral frame. In the first case the subject borrows a ready conclusion; in the second they gradually adopt the very manner of posing a moral question: which circumstances are to be considered relevant, which values are to be set against one another, in what language the conflict is to be described. It is the outsourcing of the moral framework that carries particular cultural significance, because its consequences accumulate: a single borrowed decision may remain an episode, whereas repeated recourse is capable of forming a stable habit of seeing similar situations in a particular way.

Yet not every request addressed to a system on a question with a moral dimension constitutes outsourcing. A request to establish the factual circumstances is cognitive in character; a request as to how a given situation is assessed within different moral traditions may be educational and may, on the contrary, broaden the subject's moral vocabulary. The boundary lies where the object of the request becomes the question of how this particular subject is to act in this particular situation. It is here that an external answer

may begin to perform the function ordinarily performed by one's own moral judgment.

Given the mobility of this boundary, it is worth defining operational criteria by which a particular act of interaction may be classified as outsourcing. The first is the form of the request: a question about the proper course of action for a particular subject in a particular situation is to be distinguished from a request about factual circumstances or possible modes of reasoning. The second is the presence of one's own position prior to the request: if the subject has formed no preliminary variant of an answer, it is difficult to establish to what extent the system supported, altered or in fact replaced their judgment. The third is the character of reproduction: is the subject able, after the interaction, to set out the justification of the decision in their own words without reproducing the system's ready argumentation? The fourth is the ability to name a weighty argument against the decision taken, which may serve as indirect evidence that the grounds of that decision were indeed tested by the subject.

None of these criteria is sufficient on its own. Taken together they yield a graduated scale rather than a clear dichotomy, which corresponds to the nature of the phenomenon, since outsourcing is a matter of degree rather than an event. Finally, in the broad sense it is not new. People have always sought moral advice from elders, friends, priests or literary texts. What is new are four characteristics of the present instance: round-the-clock availability with no social cost attached to the request; the absence of an evaluating gaze, which removes shame and thereby eliminates one of the regulators of moral life; the absence of any memory of the same subject's previous requests; and, most importantly, the absence of a shared fate with that subject.

3. Autonomy and Heteronomy of Choice

Kant's account of the will reveals what is philosophically at stake in outsourcing. The distinction between autonomy and heteronomy of the will, introduced in the Critique of Practical Reason, concerns not the content of an action but the source of its maxim [14, p. 41-48]. An action that coincides with the requirement of the moral law but is motivated by an external directive is not moral in the strict sense, since the will here is determined not by the subject's own practical reason but by an alien instance. From this

follows a conclusion counter-intuitive for the contemporary reader: correct advice accepted without one's own testing of its grounds does not make an action moral, but only makes it acceptable.

This consideration applies directly to our subject. If a language model systematically issues advice that a moral community would recognise as sound, this does not remove the problem but renders it invisible: heteronomy with correct content does not cease to be heteronomy, and is indeed more stable than heteronomy with incorrect content, because it generates no conflict that would draw attention to the fact of outsourcing itself.

A second support is provided by H. Arendt. Analysing the Eichmann trial, she described a specific variety of moral failure that reduces neither to evil intent nor to a deficit of intelligence, but consists in an incapacity to think – that is, to conduct the inner dialogue in which the subject tests their own maxims for acceptability [1, p. 62-75]. Judgment on her understanding is an activity, not a state. It requires effort and is undertaken afresh each time. Especially significant for our theme is Arendt's observation concerning the role of ready-made linguistic formulas: the use of formulaic language removes the necessity of thinking, since the formula already contains an evaluation.

A generative model is by its very nature a machine for producing well-formulated formulas. Its answer to a moral question is a coherent, balanced, stylistically impeccable text that offers a ready articulation of the situation. The risk discerned here consists not in the falsity of that articulation but in the fact that it exists, and that it is good, so that the need for one's own articulation disappears. Arendtian thinking [1, p. 62-75] as an inner dialogue with oneself is replaced by an outer dialogue with a system, which appears to be the same thing yet possesses none of its properties.

Kshanovskyi's findings give the first of these vulnerabilities a concrete form in the light of R. Kshanovskyi's study of philosophical task performance, discussed by O. Pankiv [22]. Testing how a language model handles tasks in philosophy, Kshanovskyi found that it works unpredictably, almost at random: to some questions it gives answers that could pass for student work, while to others it produces a grammatically impeccable mixture of the apposite and the inapposite, the reasons for this difference remaining unclear. The researcher's conclusion is that philosophy is not

a combinatorial repetition of what has already been said but a discovery of hidden, implicit meanings, and that one should therefore not expect a stream of new philosophical ideas from such a system.

Transferring this observation to the sphere of moral judgment yields a stronger argument than the one usually raised against an artificial moral adviser. The question concerns not an insufficient quality of the system that might be improved, but a mismatch between the type of product and the type of task. The system produces the most probable continuation, whereas a dilemma is exactly the situation in which the most probable solution has proved insufficient. It emerges that the subject seeks assistance in a matter for which the manner of the instrument's functioning creates a substantial limitation: the more accurately the system reproduces the most common way of resolving similar situations, the less suited it is to the one in which that way has already proved inadequate.

A third support is constituted by the postphenomenological tradition. D. Ihde has shown that technical artefacts are not neutral mediators but structure the perception of the world, illuminating some of its aspects and obscuring others [13, p. 72-80]. P.-P. Verbeek developed this thesis into the notion of moral mediation: technology participates in the formation of moral situations by determining what appears as an object of choice at all [33, p. 41-57]. B. Latour described an analogous mechanism as the delegation of action to artefacts, in which a normative requirement is built into the construction of a thing and ceases to be perceived as a requirement [20, p. 232-241]. L. Winner had earlier formulated the thesis of the politics of artefacts, according to which technical decisions embody the value priorities of their creators [36, p. 127-133].

Against this argument stands the opposing position, formulated most consistently by A. Giubilini and J. Savulescu in their conception of the artificial moral adviser. The authors proceed from the premise that human moral judgment is systematically distorted by limitations of attention, emotional bias and the selectivity of memory, and that an external system capable of taking more relevant circumstances into account and not subject to these distortions therefore brings the subject closer to the standpoint of an ideal observer and thereby improves rather than worsens their moral decisions [9, p. 174-182]. The authors stress that the adviser must remain advisory and must not make decisions on the subject's behalf.

This position deserves serious consideration, yet it contains a number of vulnerabilities. First, it treats moral judgment as a task with an optimal solution, whereas a considerable part of morally complex situations has the structure of a conflict of obligations, in which by definition no optimal solution exists. Second, on the empirical side, the conception presumes that the system is free of systematic biases, whereas the data presented below indicate the opposite. The principal objection, however, is structural and concerns not the quality of the advice but its consequences for the capacity to judge.

From the totality of these propositions it follows that a language model is neither an adviser nor an instrument in the ordinary sense. It is a mediator that structures the moral situation before the subject begins to deliberate upon it, by the very fact of the language in which it proposes that the situation be described.

4. Psychological Mechanisms of Trust in Advice

Philosophical analysis explains why moral outsourcing is a problem, but does not in itself answer the question of why the subject resorts to it. That question moves the analysis into the field of psychology. Attention should be drawn to several mechanisms that help to explain why an external automated recommendation may acquire persuasive force for a person. The first mechanism is automation bias, investigated by L. Skitka and co-authors on the material of operators of technical systems. It indicates that people tend to prefer the prompt of an automated system to their own judgment even when the available information contradicts that prompt [28, p. 998-1003]. The nature of this mechanism is twofold, ranging from cognitive economy to the transfer of responsibility onto the system.

A second mechanism consists of a pair of opposing tendencies – algorithm appreciation and algorithm aversion. J. Logg and co-authors noted that under a range of conditions people prefer an algorithmic assessment to a human one, considering it unbiased [21, p. 92-99]. B. Dietvorst and colleagues recorded the opposite tendency: having detected an error in an algorithm, a person rejects its recommendations more decisively than they would reject the recommendations of a human being who had erred in the same way [7, p. 118-123]. Both effects coexist, and their balance depends on the domain of application. The moral domain is distinctive in

that declared distrust of algorithms is highest here, yet, as the experiments discussed below show, it has little bearing on the actual persuasiveness of the advice.

A further mechanism is directly connected with the nature of moral judgment. J. Haidt's social intuitionist model has cast doubt on the assumption that moral judgment ordinarily arises as the outcome of consecutive rational reasoning. On this model, a moral evaluation often arises as a rapid intuitive reaction, while an elaborated justification is formed after the judgment has appeared and may serve the social function of explaining and defending one's own position [10, p. 814-834]. For our problematics what matters is not the opposition of intuition and rationality as such, but the distinction between the production of a moral judgment and its subsequent justification. It is at this boundary that a generative model may acquire a particular role: a user who already holds an intuitive assessment of a situation may turn to the system for a linguistically polished, coherent and persuasive justification. In such a case the system performs not so much the function of a moral adviser as that of a generator of rationalisations for a position already held.

Worth noting, finally, is the mechanism connected with A. Bandura's theory of moral disengagement. He showed that people employ devices allowing them to act against their own norms without experiencing guilt, in particular the displacement of responsibility onto an external instance and its diffusion among participants [3, p. 196-204]. Turning to a system before a difficult decision furnishes material for both devices: the decision can always be presented as having rested on external expertise.

To these mechanisms it is worth adding one further property of contemporary generative systems – a tendency towards excessive agreement with the user's position, often described in current discussion as sycophancy. For moral consultation this property is of particular significance. A system that does not call the user's initial position into question but reinforces it with well-formulated arguments may perform the function not so much of advice as of legitimation. The effect may be strengthened by the fact that the request itself frequently contains a preliminary interpretation of the situation and the desired direction of the answer. Even so, this mechanism requires separate empirical substantiation and should not be automatically ascribed to all generative models.

Developmental psychology also merits attention here, since what matters is not only the content of the norms acquired but also the subject's capacity to interpret and justify moral decisions [15, p. 347-480]. This model is of interest primarily as a ground for raising the question of the role of the experience of moral difficulty. If the transition between stages is conditioned by encounters with situations in which the available type of justification proves insufficient, then the systematic availability of a ready external justification removes specifically the deficit that drives development. This hypothesis should be formulated with caution, since it requires longitudinal verification. It is particularly important here to distinguish the adult user, whose capacity for moral reasoning is already formed, from the adolescent, who is in the process of developing it: the consequences of regular outsourcing for them are evidently different.

Finally, the historical perspective warns against any assumption that the conclusions drawn are complete. As early as 1976, describing users' reactions to the primitive program ELIZA that he had created, J. Weizenbaum recorded people's readiness to ascribe to an automated system an understanding and a competence that it did not possess [35, p. 188-201]. His conclusion was that there exist decisions that ought not to be transferred to computing systems – not because they would fail to cope, but because the transfer itself is a loss. That conclusion remains perhaps the most precise formulation of the problem with which we are dealing today.

5. Empirical Evidence of the Influence of Models

The period from 2023 to 2026 has been marked by a transition from theoretical assumptions about the possible influence of generative artificial intelligence on moral judgment to the accumulation of direct empirical evidence. Of particular importance are studies in which participants first form their own judgment, then receive a recommendation from a language model, and evaluate the same situation again. Such a procedure makes it possible to record not merely the perception of the system but an actual shift in human judgment.

One of the first systematic studies of this kind was the work of S. Krügel, A. Ostermaier and M. Uhl [17]. The authors conducted an experiment with 1,851 participants from the United States, randomly assigned to 24 experimental conditions. For the main analysis 767 participants (41%)

were retained, namely those who correctly identified both the source of the advice and its content. Participants assessed the morality of a decision in variants of the classic dilemma concerning the rescue of five people at the cost of one life, after which they received advice from a human being or from ChatGPT and repeated their assessment. The result proved fundamentally important: the advice shifted moral judgments in its own direction to a statistically significant degree. In the bridge variant the advice could even reverse the initial majority. When the adviser recommended sacrificing one person, 72% of participants considered such a decision right, whereas when the recommendation was against doing so the figure was only 17%. For ChatGPT the corresponding figures were 63% and 25%. Disclosure that the advice came from a chatbot did not eliminate the effect.

Especially telling is the gap between actual influence and the subjective sense of autonomy. Asked whether they would have made the same decision without the advice received, 80% of participants answered in the affirmative, yet their hypothetical answers continued to reproduce the direction of the recommendation's influence. The experiment recorded not only a change in moral assessment but also the inability of some participants to recognise the extent of that influence. It is precisely this discrepancy between the actual dependence of judgment and the experience of its autonomy that bears directly on the notion of outsourcing. It is correct here to speak not of a literal loss of authorship but of an experimentally detected influence and its underestimation by the participants themselves.

Another aspect of the problem concerns not the actual shift of judgment but the degree to which the moral advice of a language model appears persuasive. D. Dillion and co-authors investigated this question in an experiment [8] devoted to complex everyday moral dilemmas. Nine hundred participants evaluated 50 questions from The Ethicist column, comparing the original answers of the professional philosopher Kwame Anthony Appiah with the answers of GPT-4o. All assessments were made on a seven-point scale. GPT-4o received higher mean ratings for morality (5.63 against 5.21), trustworthiness (5.43 against 5.21), thoughtfulness (6.11 against 5.88) and correctness (5.60 against 5.40). The difference in the nuance rating was not statistically significant. On individual questions the advantage was still more apparent: the morality of the GPT answer was rated higher than that of The Ethicist in 37 of 50 cases, that is, in 74% of

situations. These results require qualification. They do not demonstrate that the model is morally more competent than a professional ethicist. The study measured the perceived quality of the advice rather than its soundness, and participants were moreover assessing answers to other people's dilemmas rather than their own. The result does show, however, that a generative model is capable of producing a text which, by socio-psychological markers of competence - elaboration, balance, stylistic confidence – is perceived as no less compelling than the advice of a recognised specialist.

A further question important for our problematics concerns the relation between trust in the source and the actual persuasiveness of the advice. E. Landes, K. Francis and J. Everett investigated it in three experiments [19]. In the first, participants received advice on everyday moral dilemmas presented as recommendations from a human being or from an LLM. Although the human adviser was rated as more reliable and more trusted, participants changed their moral judgments under the influence of the LLM to no lesser degree than under the influence of a human being. In the second study, which brought the procedure closer to real interaction with a language model, prior information about the model's performance and its assessed reliability did not determine the strength of subsequent persuasion. In the third study the quality of argumentation likewise produced no simple linear effect: good arguments did not increase persuasiveness compared with the absence of arguments, whereas manifestly absurd arguments could reduce it. This result made it possible to refine the mechanism underlying outsourcing. A person may not consider the system particularly reliable and yet still act on its advice. Psychological influence, therefore, need not pass through a settled general trust in artificial intelligence. It may be realised at the level of a particular answer, through the simple judgment that the option offered “seems good enough”. Here the fundamental boundary runs between being persuaded by an argument and situationally accepting a ready decision. The former potentially preserves the subject's autonomy, whereas the latter may contribute to its gradual replacement.

A further direction of empirical research concerns not the behaviour of users but the properties of the systems themselves. V. Cheung, M. Maier and F. Lieder have shown that large language models display cognitive biases in moral decisions that are systematically amplified in comparison with human beings: a bias towards inaction, and an effect whereby a slight

change in the wording of a question leads the system to the opposite decision. Especially important is the authors' conclusion that a substantial part of these biases appears to be introduced or amplified during post-training and alignment for dialogue interaction [5]. The moral profile of a system is thus not explained solely by the reflection of biases present in the training texts; it is also formed by subsequent engineering procedures. For the problem of outsourcing this is of principal importance. A user who perceives the system as an external source of moral evaluation has no access to the origin of the normative dispositions that determine its answer, and therefore cannot identify the persuasiveness of the advice with its moral soundness.

P. Sachdeva and T. van Nuenen approach the same picture from another side [24]. The authors analysed more than 10,000 real everyday moral dilemmas drawn from the online community Am I the Asshole (AITA) and presented them to seven large language models, comparing the resulting assessments with the decisions of Reddit users and with one another. The result revealed two characteristics that at first sight appear contradictory: the models displayed moderately high internal consistency but low agreement among themselves. In other words, an individual model could reproduce its own scheme of moral evaluation fairly consistently, yet different models frequently arrived at different conclusions about the same situation.

This result does not permit us to speak of a single “position of artificial intelligence” on moral questions. What is at issue are the positions of particular systems, formed by different training data and alignment procedures. For a user who turns to a model as an external moral arbiter this is of central importance: the form of a confident answer conceals the fact that another model – or the same model under a different protocol of interaction – might propose a different solution.

Still, the totality of these studies does not create the impression that the empirical problem has been finally resolved. Most experiments measure reactions to hypothetical situations rather than behaviour in circumstances where the participant will actually bear the consequences. Short-term designs predominate, recording situational influence but affording no grounds for judging the cumulative effect of repeated recourse. Indicative in this respect is the more recent work of T. van Nuenen and P. Sachdeva, who showed on 2,939 AITA dilemmas that a change in the narrative perspective produced a

change in the moral verdict in 24.3% of cases, whereas superficial changes to the text had a considerably smaller effect [32]. The moral judgment of a system depends, therefore, not only on the content of the dilemma but also on the manner of its linguistic presentation.

Thus the available empirical data permit a cautious but already sufficiently definite conclusion. Language models are capable of shifting human moral judgments situationally, and users underestimate the strength of this influence. The persuasiveness of a model does not require a high declared level of trust in it. The models themselves have no single stable moral position, displaying model-specific dispositions and sensitivity to the formulation of the task. Yet none of the results cited demonstrates a long-term atrophy of moral competence or a transformation of cultural transmission. It is this boundary between the established and the hypothetical that is fundamental for the cultural analysis that follows. The available research gives grounds for speaking of a short-term change in judgments and of the high persuasiveness of algorithmically formulated advice, but does not yet give grounds for asserting that repeated outsourcing regularly leads to a loss of moral autonomy.

6. Cultural Asymmetry of Moral Mediation

Taken together, the studies cited above raise a question that extends beyond the directly measurable influence of language models on individual judgment: is the moral advice generated by a system culturally neutral? This question is especially important given that a generative model does not create moral judgment out of nothing. Its answers are formed on the basis of training data and procedures of fine-tuning and alignment, and therefore bear the imprint of the cultural, linguistic and social structure of the material from which they were derived.

The problem has at least two interrelated dimensions. The first concerns the evidence base itself, whose narrowness Henrich and co-authors have documented [12, p. 63-72]. The second dimension concerns training data, which systematically represent certain languages, regions and social groups [4, p. 613-617]. For moral outsourcing this is decisive. If a model is used as a reference source, the cultural bias of its answers remains a limited problem. If, however, it becomes a participant in the production of judgment, then together with the answer a way of seeing the situation is also acquired.

How far moral evaluations diverge across societies was tested directly by the Moral Machine experiment. E. Awad and co-authors collected millions of choices from respondents in more than two hundred countries and found that preferences group into stable regional clusters with systematic differences between them [2, p. 61-63]. For our problematics this result matters not as proof of the superiority of one moral system over another, but as an argument against the assumption that there exists a single, culturally neutral form of moral intuition which a generative system might simply reproduce.

Cultural asymmetry in moral mediation should not, therefore, be understood as the simple presence of "alien values" in an answer. Considerably more complex is the situation in which a model offers the user a particular hierarchy of the morally significant without marking it as culturally conditioned. At the level of content the user may agree or disagree with a specific piece of advice. At the level of the frame they may not even notice that a particular way of posing the problem already presupposes a choice between particular values. It is this second level that appears the most significant for cultural analysis, since a frame may be acquired not through a single acceptance of an individual judgment but through a repeated practice of interaction.

In this sense it is important to distinguish the influence of an answer from the influence of the way a question is posed. The former can be measured relatively directly through the change of judgment before and after interaction with a model. The latter is considerably harder to record, since it concerns how the very set of categories through which the subject subsequently interprets moral situations is altered. If a user regularly turns to a system not merely for an argument but for a ready moral frame, outsourcing may gradually become not a single act but a way of organising their own moral experience.

In the Ukrainian context of wartime the problem acquires particular significance. This concerns not an assumption but an actual situation: the moral dilemmas widely encountered in Ukrainian society – the limits of personal responsibility, forced choice, obligation to the community, memory and forgiveness - carry a high social stake and are at the same time insufficiently represented in global language data. It is this that makes the Ukrainian material methodologically valuable: most experimental research

is built on everyday dilemmas of low stake, whereas here the situations are ones in which a moral error has a real cost.

It is important, however, not to assert that the corresponding Ukrainian dilemmas are entirely absent from training data. It is more accurate to speak of the unevenness of their representation and of the character of that representation. Part of such material may be present in the English-language digital space, but it often enters it through journalistic, academic or political texts produced by external observers. The internal moral speech of the community – everyday disputes, private arguments, local contexts, linguistic nuances, informal ways of passing on experience – is potentially represented differently. This is this difference that requires separate empirical verification.

In such a context the linguistic dimension of the problem acquires particular significance. Ukrainian is not merely a channel for conveying an already formed moral content. It is the medium in which moral problems receive their names, their distinctions and their argumentation. The question whether a model's moral answer differs depending on the language of the request therefore does not reduce to stylistic or translational differences. Potentially at issue is a difference in the set of arguments and moral categories that the system makes available to the user. A comparative test of answers to identical dilemmas in Ukrainian and in English could show whether such an asymmetry exists in fact. At the present stage this should be regarded as a research hypothesis.

No less important is the question of the type of moral frame that a system offers under conditions of war. The situations arising in such experience have for the most part the structure of a conflict of loyalties rather than a conflict of interests. Obligation to one's family collides with obligation to the community, the preservation of life with obligation to the country. Resolving such a dilemma within one's own tradition presupposes not a weighing of consequences but a determination of which loyalty is the higher, and an acceptance of guilt for violating the lower. The problem is especially pronounced where a moral decision requires not the immediate relief of tension but the capacity to hold a question open. Some Ukrainian dilemmas have no resolution in the present tense, whereas the instrument by its very construction issues an answer immediately.

This yields a distinction fundamental for cultural analysis. Moral outsourcing may alter not only individual assessments but also the moral framework itself, that is, the way moral questions are posed – which circumstances are considered relevant, which values are set against one another, in what language the conflict is described. At this point the problem of cultural asymmetry converges with the problem of value transmission: cultural transfer occurs not only through the content of convictions but also through the ways in which they are produced and justified.

7. A Typology of the Risks of Outsourcing Judgment

On the basis of the analysis conducted it is worth setting out the risks by corresponding levels, from the individual to the socio-institutional. The distinction of levels is essential, since the countermeasures appropriate to each differ. The individual level covers the consequences for the particular subject. The leading risk here is moral deskilling, a notion drawn by S. Vallor from virtue ethics: moral competence is partly constituted by practical skills acquired through repeated exercise, which may weaken when the occasions for such exercise diminish, and technologies that take over the performance of morally loaded actions reduce those occasions [30, p. 110-118]. To this are added the illusion of soundness, in which the fluency of a text is taken for depth of reasoning; heteronomy under the guise of autonomy; and the formation of dependence, in which recourse to a system replaces recourse to people.

At the epistemic level what is affected are the conditions under which a person can know the morally significant at all. One of the principal risks here is the attribution of unwarranted authority to the system. It has no body, no interest of its own and no biography, and the user is inclined to read this absence as impartiality, whereas in reality it means only that the source of the position remains opaque.

Weighty risks are present at the cultural level. These are, in the first place, the asymmetry with which different moral traditions are represented, and the displacement of local moral vocabularies, that is, of the linguistic categories in which a community names and distinguishes moral situations. Next is the weakening of the channel of intergenerational transmission: moral advice is increasingly sought from an instance that does not share the community's lived history or collective memory and bears no part in

its future. There is also the loss of the dilemma as a cultural form: a tragic choice, which within a tradition exists as insoluble and for that very reason gives rise to tragedy, to the novel and to moral reflection, is reformatted into a task with an optimal solution. S. Turkle has described a mechanism structurally identical to the argument about deskilling: the displacement of conversation by the exchange of messages reduces the number of situations in which a person must endure the discomfort of direct contact, and thereby diminishes the capacity for empathy, which is formed only in such situations [29, p. 21-35]. Seeking moral advice from another person has always been a socially costly act, and those costs performed a function, embedding the request within the fabric of relationships. Recourse to a system is costless and therefore generates no relationship.

At the socio-institutional level the consequences fall on social structures. The leading risk is a rupture of responsibility. In a chain where a person asked, a system answered and a person acted, there is no point at which responsibility can be unambiguously located. Next is the erosion of moral communities as sites of shared deliberation, since individual recourse removes the need for collective discussion. Inequality of access to high-quality systems constitutes a new dimension of cultural inequality. Finally, there is political vulnerability, since an infrastructure to which people turn en masse for moral orientation is an attractive target for influence.

One further consideration takes the problem beyond a simple enumeration. First, it is necessary to distinguish notions that are frequently conflated in discussions of morality and technology. Moral evil in the classical understanding presupposes a subject, an intention and responsibility. The banality of evil in Arendt's sense describes harm inflicted without evil intent, yet by a person who has refused to think. Artificial evil is a special term denoting moral harm arising in information systems as a consequence of imperfect design or of misuse. Cumulative technologically mediated harm is a broader category, encompassing the consequences of an aggregate of conscientious but structurally uncoordinated actions. Moral outsourcing is described most accurately by this last category.

The structure of the process is transparent enough. A developer configures the system to be cautious, guided by entirely good considerations. The user turns to it wishing to act rightly. The system produces a balanced, correct answer. Yet the aggregate consequence of millions of such interactions

may be something none of the participants desired: a narrowing of the community's moral vocabulary and a weakening of its members' capacity to judge independently.

It is here, it would seem, that the boundary runs between the ethics of technology and cultural policy. The ethics of technology operates with the categories of intention, responsibility and accountability, and therefore works poorly where harm is the cumulative effect of conscientious actions. Cultural policy, by contrast, is accustomed to dealing exactly with such processes: the decline of a language, the loss of an activity, the disappearance of a tradition rarely have a culprit, yet they are entirely amenable to analysis and, to a degree, to regulation. From this it follows that instruments for resolving the problem described are to be sought rather in the second of these spheres.

8. Conditions for Productive Interaction with a Model

An analysis that ends merely with a list of risks is incomplete. The available data give grounds for the conclusion that the influence of language models on moral judgment is not uniform and depends on the mode of interaction. The distinction proposed by E. Landes and co-authors indicates that between a change of position under the influence of an argument and a change of position through submission to an instance there are formed the conditions under which interaction sustains moral competence [19].

The first condition is the sequence of steps. Interaction works if the subject formulates a position before turning to the system, and fails if they turn to it instead of formulating one. The technically simple requirement of recording one's own judgment in advance changes the character of the subsequent interaction fundamentally: the presence of a fixed position turns the system's answer into an object of comparison rather than a source.

The second condition is the mode of interaction. A request for arguments against the subject's own position is productive; a request for the answer is not. A system asked to find the strongest objections performs the function that in the moral tradition was performed by an opponent, and in doing so does not replace the act of choice. A request for a conclusion, by contrast, is outsourcing in its pure form.

The third condition is the plurality of sources. Since agreement between different systems is low, putting the same question to several of them

reveals a divergence that is itself valuable evidence: it restores to the user the understanding that what is at issue is a position and not a truth.

The fourth condition is educational. It echoes the thesis advanced in the literature concerning the need to form a digital culture in which users are aware that they can no longer reliably establish the origin of a text and must therefore be more critical [22]. In the sphere of moral judgment this precept takes a concrete form: criticality should be directed not at the verification of facts, which in a moral situation are rarely in dispute, but at recognising the value frame within which the answer has been formulated. Hence a concrete task for the educational system - the formation of the skill of recognising the moment at which a justification has ceased to be one's own. This is a skill and not a body of knowledge, and it requires exercise. A practical form might be work on the analysis of dilemmas in which a student consecutively performs three steps: formulates a position, receives a generated alternative, and justifies agreement or disagreement with it. Such a format does not prohibit the use of the instrument but embeds it in a structure that preserves exercise.

The fifth condition concerns developers and regulators. Since a substantial part of the systems' moral biases appears to originate in post-training and alignment [5], the requirement of transparency regarding these stages is not abstract but operational. A user who seeks moral advice has the right to know that the system's caution is an engineering decision rather than a moral position, and that it systematically shifts recommendations towards abstention from action.

CONCLUSIONS

Moral outsourcing is an independent phenomenon, distinct from moral consultation, moral enhancement and moral automation. Its specificity lies in the rupture between the formally retained authorship of a decision and the actual transfer of the work of judgment. Culturally more significant is the outsourcing not of an individual decision but of a moral frame.

The correctness of the content of advice does not guarantee the autonomy of a decision. Within the Kantian framework, a judgment repeated without one's own testing of its grounds acquires the features of heteronomous acceptance. Arendt's understanding of judgment as an activity refines the mechanism, while the postphenomenological tradition

adds the understanding that a system structures a moral situation before deliberation upon it has begun.

The persuasiveness of algorithmic advice is conditioned by an aggregate of mechanisms: automation bias, algorithm appreciation, the transfer of responsibility, the correspondence between the system's product and the need for a ready justification of an already existing intuition, and the tendency of systems to agree with the user. The psychological theories drawn upon are used as partial explanations rather than as components of a single model.

Empirical data from 2023 to 2026 attest to a short-term influence of language models on moral judgment under controlled conditions. Some systematic biases appear to be introduced or amplified during post-training and alignment, and remain largely opaque to users.

The risks of outsourcing manifest themselves at the individual, epistemic, cultural and socio-institutional levels. The leading risks are moral deskilling, the attribution of unwarranted authority to the system, the possible displacement of local moral vocabularies and a rupture of responsibility. None of these risks is inevitable: each is realised under particular conditions of interaction.

The moral harm resulting from widespread outsourcing is described as cumulative technologically mediated harm. It is the consequence of conscientious but structurally uncoordinated actions on the part of all participants and has no author. The proper field for addressing it is cultural policy and education rather than an ethics of technology oriented towards intention and accountability.

Interaction with a language model can sustain moral competence provided that the subject formulates a position in advance, requests objections rather than conclusions, consults several systems, engages with educational formats that preserve exercise, and that developers are transparent about post-training and alignment. Prospects for further research are connected above all with the empirical verification of the assumptions advanced here.

References:

1. Arendt H. (2013). Banalnist zla. Sud nad Aikhmanom v Yerusalymi [Eichmann in Jerusalem: A Report on the Banality of Evil] (A. Kotenko, Trans.). Kyiv: Dukh i Litera, 367 p. (in Ukrainian)

2. Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I. (2018). The Moral Machine experiment. *Nature*, vol. 563, pp. 59–64. DOI: 10.1038/s41586-018-0637-6

3. Bandura A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, vol. 3, no. 3, pp. 193–209. DOI: 10.1207/s15327957pspr0303_3

4. Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, pp. 610–623. DOI: 10.1145/3442188.3445922

5. Cheung V., Maier M., Lieder F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, vol. 122, no. 25, article e2412015122. DOI: 10.1073/pnas.2412015122

6. Chowdhury R. (2018). Moral outsourcing: Humanity in the age of AI. TEDxOccidentalCollege, June 8, 2018. Available at: https://www.ted.com/talks/rumman_chowdhury_moral_outsourcing_humanity_in_the_age_of_ai

7. Dietvorst B. J., Simmons J. P., Massey C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General,* vol. 144, no. 1, pp. 114–126. DOI: 10.1037/xge0000033

8. Dillion D., Mondal D., Tandon N., Gray K. (2025). AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, vol. 15, article 4084. DOI: 10.1038/s41598-025-86510-0

9. Giubilini A., Savulescu J. (2018). The artificial moral advisor: The "ideal observer" meets artificial intelligence. *Philosophy & Technology*, vol. 31, no. 2, pp. 169–188. DOI: 10.1007/s13347-017-0285-z

10. Haidt J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, vol. 108, no. 4, pp. 814–834. DOI: 10.1037/0033-295X.108.4.814

11. Hassen M. Z. (2025). The risk of moral outsourcing: Why artificial intelligence cannot and should not make our ethical decisions. *EthAIca*, vol. 4, article 428. DOI: 10.56294/ai2025428

12. Henrich J., Heine S. J., Norenzayan A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, vol. 33, no. 2–3, pp. 61–83. DOI: 10.1017/S0140525X0999152X

13. Ihde D. (1990). Technology and the Lifeworld: From Garden to Earth. Bloomington: Indiana University Press, 226 p.

14. Kant I. (2004). Krytyka praktychnoho rozumu [Critique of Practical Reason] (I. Burkovskyi, Trans.). Kyiv: Yunivers, 240 p. (in Ukrainian)

15. 15. Kohlberg L. (1981). Essays on Moral Development. Vol. 1: The Philosophy of Moral Development. San Francisco: Harper & Row, 441 p.

16. Kontseptsiia rozvytku shtuchnoho intelektu v Ukraini [The Concept of the Development of Artificial Intelligence in Ukraine] (2020). Approved by the Order of the Cabinet of Ministers of Ukraine of December 2, 2020, no. 1556-r. Available at: https://zakon.rada.gov.ua/laws/show/1556-2020-%D1%80 (in Ukrainian)

17. Krügel S., Ostermaier A., Uhl M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, vol. 13, article 4569. DOI: 10.1038/s41598-023-31341-0

18. Kryvonos O., Kozlovets M. (2026). Vykorystannia heneratyvnoho shtuchnoho intelektu pid chas intelektualnykh zmahan ta konkursiv: etychni aspekty [The use of generative artificial intelligence in intellectual competitions: ethical aspects]. *Humanities Studies*, vol. 26 (103), pp. 47–54. DOI: 10.32782/hst-2026-26-103-05 (in Ukrainian)

19. Landes E., Francis K. B., Everett J. A. C. (2026). People defer to AI moral advice, but not blindly. Cognition. DOI: 10.1016/j.cognition.2026.106215

20. Latour B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. Shaping Technology / Building Society: Studies in Sociotechnical Change / eds. W. E. Bijker, J. Law. Cambridge, MA: MIT Press, pp. 225–258.

21. Logg J. M., Minson J. A., Moore D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90–103. DOI: 10.1016/j.obhdp.2018.12.005

22. Pankiv O. V. (2023). Etychni pytannia vykorystannia tekhnolohii mashynnoho navchannia [Ethical issues of the use of machine learning technologies]. *Visnyk Lvivskoho universytetu. Seriia filosofsko-politolohichni studii*, vol. 46, pp. 147–153. (in Ukrainian)

23. Rekomendatsiia pro etychni aspekty shtuchnoho intelektu [Recommendation on the Ethics of Artificial Intelligence] (2021). Adopted by the General Conference of UNESCO on November 23, 2021. Paris: UNESCO, 44 p. (in Ukrainian)

24. Sachdeva P., van Nuenen T. (2025). Normative evaluation of large language models with everyday moral dilemmas. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM. DOI: 10.1145/3715275.3732044

25. Schönpflug U. (2001). Intergenerational transmission of values: The role of transmission belts. *Journal of Cross-Cultural Psychology*, vol. 32, no. 2, pp. 174–185. DOI: 10.1177/0022022101032002005

26. Shevel I. (2025). Kultura shtuchnoho intelektu, chat-botiv ta emodzi yak novyi etap kulturnoi komunikatsii v tsyfrovu epokhu [The culture of artificial intelligence, chatbots and emoji as a new stage of cultural communication in the digital age]. *Pytannia kulturolohii*, vol. 45. (in Ukrainian)

27. Shvets I. (2025). Tsyfrova inkliuziia v kulturolohichnomu dyskursi [Digital inclusion in the discourse of cultural studies]. *Pytannia kulturolohii*, vol. 45. (in Ukrainian)

28. Skitka L. J., Mosier K. L., Burdick M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, vol. 51, no. 5, pp. 991–1006. DOI: 10.1006/ijhc.1999.0252

29. Turkle S. (2015). Reclaiming Conversation: The Power of Talk in a Digital Age. New York: Penguin Press, 436 p.

30. Vallor S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, vol. 28, no. 1, pp. 107–124. DOI: 10.1007/s13347-014-0156-9

31. Vallor S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford: Oxford University Press, 309 p.

32. van Nuenen T., Sachdeva P. S. (2026). The fragility of moral judgment in large language models. arXiv preprint arXiv:2603.05651. DOI: 10.48550/arXiv.2603.05651

33. Verbeek P.-P. (2011). Moralizing Technology: Understanding and Designing the Morality of Things. Chicago: University of Chicago Press, 183 p.

34. Voronkova V. H., Nikitenko V. O. (2022). Filosofiia tsyfrovoi liudyny i tsyfrovoho suspilstva: teoriia i praktyka [The Philosophy of the Digital Human Being and Digital Society: Theory and Practice]. Lviv–Torun: Liha-Pres, 460 p. DOI: 10.36059/978-966-397-250-3 (in Ukrainian)

35. Weizenbaum J. (1976). Computer Power and Human Reason: From Judgment to Calculation. San Francisco: W. H. Freeman, 300 p.

36. Winner L. (1980). Do artifacts have politics? *Daedalus*, vol. 1*09, no. 1, pp. 121–136.

37. World Commission on Environment and Development (1987). Our Common Future. Oxford: Oxford University Press, 400 p.