Strengthening the AI Operating Environment

Distributed competence as a means to risk mitigation

Bruce Hedin[1], Samuel Curtis[2]

[1] *Hedin B Consulting*
[2] *The Future Society*

Abstract
In the rapidly evolving discourse on artificial intelligence (AI), the familiar refrain of "maximizing potential while mitigating risks" has become somewhat of a ubiquitous mantra, emphasizing the need for an effective risk mitigation framework. This paper briefly examines the current state of AI-enabled applications and discusses the various risk containment strategies being implemented. Initial efforts focused on establishing high-level principles for responsible AI use. More recent strategies have sought to operationalize these principles through normative instruments, such as industry best practices and legal statutes, that govern AI applications and their creators. While valuable, such a top-down approach is not sufficiently effective; a complementary, bottom-up approach focused on strengthening the environment in which AI is deployed is also necessary. The paper analyzes two specific initiatives aimed at enhancing the human component of AI deployment (creating a better-informed public through AI benchmarks, creating a better-equipped public with resources for local validation) and offers insights on how this environment-focused track can contribute to risk containment. Furthermore, we suggest additional steps for leveraging this approach in tandem with top-down strategies to cultivate a more robust risk mitigation framework.

Keywords
AI governance, AI education, AI risk, Benchmarking, Evaluation, Validation, Effectiveness, Competence, Trust

1. Introduction

As the use of AI-enabled applications, both in the legal domain and elsewhere, has gone from a topic for academic discussion to a matter of everyday practice, questions about how best to realize the potential of such applications, and how best to mitigate the risks attendant upon their use, have taken front and center in the various venues in which the interaction between technology and the norms and institutions that govern the life of society are discussed. This attention to AI's potential for both good and bad, and to ways of realizing the former while containing the latter, has only been heightened in recent months by the release of a range of publicly accessible applications that draw on large language models (such as GPT-4).

An attention to risks attendant on the use of AI, provided it is grounded in an understanding of AI's real capabilities and limitations, is salutary. It is true that the risks, given the current state of the technology, are sometimes overstated (LLMs are indeed robust platforms for a range of different applications and can generate output that closely approximates that which a human might create; they are, nevertheless, still simply statistical models of discourse tokens, well short of the capacity for understanding and creativity characteristic of general intelligence[1][2]). It is also true, however, that even narrow-purpose AI applications (e.g., within the legal domain, those designed specifically for judicial decision modeling, predictive policing, or facial recognition) can, if used improperly, jeopardize core social values such as fairness, subtract from individual privacy, liberty, and dignity, and undermine assumptions about truth-seeking and justice-realization that are the basis for the rule of law (and hence for a stable democratic order). These are, regardless of one's perspective on the capacity and implications of LLMs, serious risks that call for commensurate efforts at risk containment.

Efforts at containing the risks attendant upon AI have been under way for some time. Early efforts focused on articulating high-level, value-oriented, principles for the responsible design, development, and use of AI (for examples, see: [3][4][5][6][7][8]). Collectively, these efforts were, if the sheer quantity of principles (or sets of principles) proposed is a measure of success, quite successful[9][10][11]. Where these efforts fell short was in establishing mechanisms connecting the principles to actual practice.

More recent efforts, seeking to fill this gap, have focused on the question of how to operationalize such principles. The objective of these efforts has generally been the creation of normative instruments that would encourage, or enforce, adherence to the aspirational principles. The forms proposed for such normative instruments have

In: *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023),* held in conjunction with ICAIL 2023, June 19, 2023, Braga, Portugal.

bhedin@hedinb.com (B. Hedin);
samuel.curtis@thefuturesociety.org (S. Curtis)
https://thefuturesociety.org/ (S. Curtis)

varied, from informal industry best practices, to more precise (and auditable) standards, all the way to enforceable legal statutes. The object of governance for these normative instruments has been primarily AI applications and their creators: the norms established are intended to act as "guardrails" on the design and operation of AI-enabled applications, on the objectives and requirements of developers of applications, on the use cases in which applications may be deployed, and even on the structure and conduct of the entities that produce AI-enabled applications. Notable examples of initiatives on this *top-down* track include the creation of government offices charged with responsibility for algorithm inspection, proposals for regulations requiring that AI applications meet certain design specifications ("privacy by design," "human rights by design"), laws requiring the destruction of training data, restrictions or outright bans on the use of certain applications (judicial modeling technologies, facial recognition technologies), and calls for a global moratorium on the research and development of "strong" AI.

This top-down, application-focused, approach to risk containment is, in at least some of its less heavy-handed instantiations, a valuable and necessary one. It does not, however, exhaust the approaches to risk containment available to policymakers and other stakeholders in the safe use of AI. Complementary to the application-focused approach is an approach that starts from a *bottom-up* perspective and takes as its objective, not the creation of guardrails on the development and use of AI, but rather the strengthening (or "hardening") of the environment in which AI-enabled applications are deployed. This approach seeks to contain risk by making the environment (in all its components: hardware, software, and human) in which AI is deployed more resistant to AI misuse (whether intentional or not) and therefore less susceptible to the risks attendant on such misuse.

In this paper, we examine more closely the potential that the bottom-up, environment-focused, track holds as a means for risk containment. We do so by considering approaches to strengthening the human component of the environment in which AI-enabled applications are deployed. More specifically, we draw attention to two key gaps in the resources currently available to stakeholders in the responsible use of AI in the service of the law: (1) the absence (discussed in Section 3) of an on-going program of benchmarks that can provide stakeholders with meaningful information on the actual capabilities and limitations of AI-enabled legal applications and (2) the absence (discussed in Section 4) of resources that would allow practitioners to conduct their own evaluations of the effectiveness of AI in real-world settings. In the case of each gap, we characterize the nature of the need, identify the features of a solution that would meet the need, and discuss work done to date toward such a solution. With the perspective gained from this discussion, we draw (in Section 5) some general lessons about the potential the environment-focused track holds for risk containment.

2. Related work

This paper offers a framework (simply put: top-down vs. bottom-up) for assessing approaches to mitigating the risks attendant on the use of AI-enabled applications in the service of the law. There are, of course, other frameworks that have been offered and these also can provide insightful perspectives. Among the initiatives that are related to, and often complementary to, the work presented in this paper are the following.

Guidelines. The Asilomar AI Principles [7] put forward 23 principles spanning research issues, ethics and values, and longer-term issues, for the research and development of AI. European Ethical Charter on the Use of AI in the Judicial Systems and their Environment [5] presents five principles, intended for both public and private stakeholders responsible for the design and deployment of AI tools and services that involve the processing of judicial decisions and data, and were adopted by the European Commission for the Efficiency of Justice (CEPEJ). The General Principles of Ethically Aligned Design [4] proposes eight principles upon which ethical and values-based design, development, and implementation of autonomous and intelligent systems (including artificial intelligence and intelligent assistance technologies designed for legal professionals; see the Chapter on Law) should be guided. The European Commission's Ethics guidelines for trustworthy AI [6], drafted by the European Commission High-Level Expert Group on AI, puts forward seven key requirements that AI systems should meet in order to be deemed trustworthy. The Partnership on AI has drafted eight tenets [8] that its members, spanning industry, academia, and non-profit, "endeavor to uphold."

Risk mitigation frameworks focused specifically on LLMs. Weidinger et al.[12] proposes a comprehensive taxonomy of ethical and social risks associated with large-scale language models, identifying twenty-one risks across six risk areas and discussing approaches to risk mitigation. Bai et al.[13] proposes a method called Constitutional AI (CAI) to train a non-evasive and relatively harmless AI assistant without human feedback labels for harms, with the aim of developing techniques to create AI systems that adhere to design (or "constitutional") principles, as opposed to learning from human feedback. Mökander et al.[14] proposes a three-layered approach to auditing large language models, which includes governance audits, model audits, and application audits—the third of which includes a component to check

LLMs' adherence to ethical principles.

Educational efforts. Long and Magerko[15] provides a concrete definition of AI literacy based on existing research and synthesizes a variety of interdisciplinary literature into a set of core competencies of AI literacy, as well as design considerations to support AI developers and educators in creating learner-centered AI. Lin and Van Brummelen[16] presents the findings from workshops co-designed with K-12 teachers—that scaffolding in AI tools and curriculum is needed for ethical and data discussions, learner evaluation, engagement, peer collaboration, and critical reflection—and an exemplar lesson plan illustrating ways to teach AI in non-computing subjects within a remote setting. Gašević et al.[17] explores the theme of empowering learners for the age of AI and highlights the need for foundational discussions about learning theory and conceptualizations of learning actions and behaviors in AI-human settings, as well as concerns regarding ethics, bias, and fairness in AI's growing influence. Hugging Face[18] has sought to democratize machine learning knowledge and competence by offering educational materials for beginners as well as instructors. Hugging Face supported the BigScience open research collaboration, which brought together more than 1,000 researchers from 60 countries and more than 250 institutions to create BLOOM[19], an openly and transparently trained multilingual LLM.

3. Strengthening the AI operating environment through better information

An environment in which those who would use, or be affected by, AI-enabled applications lack at least baseline information about AI (what it is, where it is, how it works, and how *well* it works) is an environment conducive to misuse (not to mention to unhelpful, even harmful, hype). Conversely, an environment in which both active and passive users of AI are well-informed about AI's use cases, its conditions of use, and its strengths and weaknesses is one that will be more resistant to AI misuse (and to the risks associated with such misuse). An important component of any effective strategy for containing the risks associated with AI will therefore be *education*: if we can foster a public that is better informed about AI, we will foster a public better equipped to recognize and guard against the risks associated with it.

The role of education in risk containment has been recognized for some time. Education was one of the three themes of the inaugural edition (2019) of The Athens Roundtable on Artificial Intelligence and the Rule of Law[20]. Education is also the focus of a number of current initiatives. The Future Society, to cite one example, has developed a MOOC on AI and the Rule of Law[21]; aiEDU, to cite another, is an initiative that promotes broad AI literacy through the development of AI curricula for use in a wide range of educational venues, from K-12 schools to public museums[22].

As potentially valuable as these educational initiatives are, they will be successful in meeting their objectives only insofar as they are able to access and convey accurate and meaningful content. This is where a challenge appears: for some topics, namely topics related to the effectiveness of AI, the content is lacking (or at least lacking in the form required for fostering broadly distributed AI competence). In this section, we examine this gap and consider an approach to filling it.

3.1. The need

If we wish to foster an informed, and empowered,[1] public, one capable of making empirically well-grounded decisions about the sorts of tasks to which AI should and should not be applied, and the conditions that should be met when it *is* applied, we need to ensure that the public has access to accurate information about the effectiveness of AI (i.e., its capabilities and limitations when applied to real-world tasks). The problem is that evidence of effectiveness of AI-enabled applications is spotty: it exists, and is accessible, in only a very incomplete and inconsistent manner. The reason is that there is no suitably authoritative institutionalized program for generating the required evidence in a manner and format that can be readily consumed by individuals and civil society groups, thereby meeting the objective of giving citizens informed agency over the use of AI in their (and their fellow citizens') lives.

It is also worth noting that, while our focus in this section is on the means to foster an informed public, the evidence gap just observed has wider implications. It acts as roadblock not only to meeting the objective of an informed public but also to meeting the other objectives of the principles that have been articulated for the responsible use of AI (which may be stated,[2] at an abstract level, as (1) protection of core values, (2) creation of conditions needed for an informed trust, and (3) advancement technological innovation and economic prosperity). Without sound evidence of effectiveness, **we will be unable to protect core values**, because we won't know (a) whether the AI-enabled systems achieve their immediate goals nor (b) whether, even in achieving those immediate

[1] Informed, of course, does not necessarily mean empowered. Advancing the empowerment of citizens means not only ensuring that citizens have access to information but also ensuring that the legal and practical conditions are such that citizens can act on that information.

[2] This threefold classification scheme for the objectives of principles is the authors' own; other classification schemes can be found in [9] and [10]

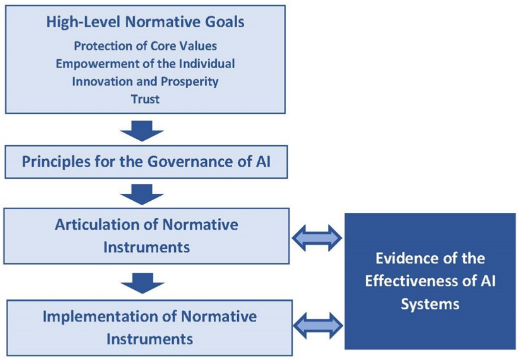
High-Level Normative Goals
Protection of Core Values
Empowerment of the Individual
Innovation and Prosperity
Trust
Principles for the Governance of AI
Articulation of Normative Instruments
Implementation of Normative Instruments
Evidence of the Effectiveness of AI Systems

Figure 1: The role of evidence of effectiveness in the cascade from aspirational values to viable normative instruments.

goals, they impinge on other core values. **We will be unable to create the conditions of trust**, because we will lack the empirical data that is the basis for a well-grounded trust (or distrust) [23]. **We will be unable to advance the goals of technological innovation and economic prosperity**, because we will lack the information needed to optimize the allocation of research effort and financing. In terms of approach, moreover, having access to sound evidence of effectiveness is necessary for both bottom-up and top-down approaches to risk containment. With regard to the latter, as illustrated in Figure 1, evidence of effectiveness is necessary both for the formulation of viable normative instruments and for the assessment of adherence to the norms instantiated in such instruments. In short, evidence of effectiveness is needed both for the general objective of ensuring the responsible use of AI-enabled systems and for the specific objective of fostering an informed public.

Now, to say that the required evidence is lacking is not to say that there is no evidence at all. There is indeed a healthy flow of reports, of various types, of evaluations of the effectiveness of AI-enabled systems. These include: academic research papers[24][25][26]; industry white papers; reports of government-sponsored evaluations[27][28]; evaluations conducted by non-governmental civic organizations[29][30][31]; and academic and industry-sponsored benchmarking initiatives[32][33].

The problem is that these evaluations, while well-designed to meet their own objectives, have not been designed specifically to meet the objective of fostering a general public that is informed and empowered. As a result, the evidence the studies generate is lacking in key features required to meet that objective. Among key limitations of current evaluations[3] are the following.

- **Narrow focus.** The research objectives of the evaluations are often such that they are better served by narrowly circumscribing the scope of the exercise, not measuring the impact of the whole sociotechnical system, of which the technology is a part, on the values with which the public may be concerned. Consistent with these objectives, the evaluations gauge the performance of the technologies being evaluated using metrics specifically relevant to the capability addressed in the study; they do not seek measures that would provide a comprehensive view of the technology's fitness for purpose.
- **Distance from real-world circumstances.** In the interest of arriving at a well-controlled answer to specific research questions, the studies often do not make allowance for variability in all the factors that could, in a real-world setting, affect a system's effectiveness. The result is an exercise that is removed from the real-world circumstances. Moreover, obtaining evaluation data sets that, in both size and character, are reflective of the data populations to which the technology under evaluation would be applied in a real-world setting is a challenge that current evaluations are often unable to meet.
- **Misalignment of purpose.** Many currently available evaluations are of the one-off variety: they are designed to produce just the data needed for the study that occasioned them and they are not intended to be repeated on a regular basis. An additional limitation that is particularly characteristic of industry white papers is that they are generally designed, not to provide a well-rounded view of the technology's fitness for purpose, but to highlight characteristics of the enterprise's offering that, the enterprise believes, will resonate in the marketplace.

3.2. A proposal for meeting the need

If the objective of an informed (and empowered) public is a worthy one, and if a lack of evidence of the effectiveness of AI-enabled systems is impeding the achievement of that goal, then what might a solution that removed that impediment look like? What we propose, and what we discuss in the remainder of this section, is the creation of an on-going institutionalized program of *interoperable open* AI benchmarks, the purpose of which would be to supply the empirical evidence needed to foster a public empowered to make informed decisions about the use of AI-enabled technologies. The benchmarks should be "open" in the sense that exercises must be transpar-

[3] We are, of course, not of saying that all currently available evaluations are subject to all of these limitations. We are saying simply that each evaluation is subject to at least one of them.

ent: data used, procedures followed, and results generated must all be open to inspection (or, in some cases, audit), by both participants and independent observers. They should be "interoperable" in the sense that they will supply evidence usable by all regulatory regimes, regardless of the specific goals and priorities that are operative within any specific jurisdiction. Furthermore, if they are to serve their intended purpose of fostering a better-informed public, they should generate results that can be understood by both experts and non-experts.

3.2.1. Requirements

A benchmarking program that will meet the general objective of fostering an informed public (a public that includes everyone from researchers and designers, to policymakers and lawyers, all the way to the potentially involuntary decision subjects of judicial or enforcement technologies) will have to meet certain requirements. It must: (1) design evaluations that model real-world circumstances; (2) generate results that will be meaningful, and actionable, for a wide range of stakeholders; (3) run evaluation exercises that are consistent and trusted; and (4) be practically viable. Specific implications of these basic requirements are the following.

- **Real-world.** In order to be relevant to real-world practice, it is essential that the evaluations conducted in the benchmarking program (1) closely model real-world conditions and objectives and (2) take as the target of their measurement the whole system of which the AI-enabled technology is a part. A failure to do so would be a failure to provide the evidence actually required by the public.
- **Meaningful.** In order to be actionable, it is essential that the results generated by the benchmarking program (1) be expressed via meaningful metrics and (2) be interoperable across national and other jurisdictional boundaries. With regard to metrics, "meaningful" means that they should be (1) statistically sound, (2) relevant, and (3) and understandable both to experts and non-experts. The interoperability requirement means that the results of the exercises should be broadly usable, providing information that can be acted upon regardless of the specific goals and priorities that are operative within any specific jurisdiction.
- **Consistent and trusted.** If the public is to rely upon the results produced by a benchmarking program, the results must be generated in a consistent and trusted manner. This means, specifically, (1) that the evaluations should be run on a periodic schedule, (2) that the evaluations should be of a reasonably consistent design (at least consistent enough to allow informative comparison from one run of an exercise to the next), (3) that the program should be institutionalized (i.e., have the legitimacy and durability that come from sponsorship by recognized public authorities), and (4) that the design and execution of the evaluations run in the program be transparent (data used, procedures followed, and results generated must all be open to inspection by both participants and independent observers).
- **Practical.** In order to be viable, the program must also meet a number of non-trivial practical requirements. These include (1) reaching consensus on metrics for concepts and tasks where that consensus is currently elusive[34], (2) obtaining fresh and meaningful data sets on a regular basis, (3) achieving broad participation (which means having low barriers to entry, in terms of both cost and reputational risk), and (4) producing its results in a timely and efficient manner.

Meeting the requirements and challenges on this list will not be a trivial undertaking. Fortunately, those who would create a benchmarking program aligned with this vision are not without resources upon which to draw. As we have already seen, researchers have been designing and conducting evaluations of AI-enabled systems for many years. While those evaluations have not been designed for the same purposes as those that would be run in the proposed benchmarking program, they can still serve as a valuable resource for those seeking to address the requirements and challenges of a meaningful AI benchmarking program. A few examples of such resources are the following.[4]

- The series of studies conducted in the NIST-sponsored Text Retrieval Conference (TREC)[27].[5]
- The HELM (Holistic Evaluation of Language Models) initiative undertaken by Stanford's Center for Research on Foundation Models [33].
- METRICS – An international competition for the evaluation of robotics and AI[36].
- NIST's 2021 AI Measurement and Evaluation Workshop[37].

[4] It is worth emphasizing that this is simply a selection of examples, not an exhaustive list of available resources.

[5] We note that the series of evaluations conducted in the TREC Legal Track from 2006 through 2011[35], which produced data on the effectiveness of advanced technologies at the task of legal discovery, and thereby provided empirical grounding for decisions as to whether to adopt those technologies (or, in the case of courts, to allow their adoption), illustrates the potential that a well-designed *on-going* program of benchmarks could hold for creating a better-informed public.

- A framework developed by the AI Ethics Impact Group for operationalizing AI ethics[38].
- Academic papers published in the proceedings of AI-focused conferences[25][26].

The above is of course not an exhaustive list of available resources; it is intended simply to illustrate the sorts of resources may build upon in developing a benchmarking program that would meet the need we have identified.

3.2.2. Benefits

While designing and implementing a program that meets the requirements we have identified would be a challenge, the benefits of meeting that challenge are significant and tangible. By filling the evidence gap, the program would help to foster a public that was better informed about the real capabilities, limitations, and risks of AI-enabled systems (including those drawing upon LLMs). It would do so both directly, insofar its results were consumed by members of the public without the mediation of other entitiies, and indirectly, insofar as its results reached the public through the mediation of civil society groups or educational initiatives focused on questions of society and technology. A better informed public would, in turn, be one better positioned to recognize, and address, risks to core human values, to protect the liberty, privacy, and dignity of the individual, to resist the temptation of unwarranted hopes or fears about AI, and to support measures that further, in a responsible manner, scientific innovation and economic prosperity.

Apart from these primary benefits, such a program would also bring a number of collateral benefits. These include: (1) thanks to its provision of empirically sound and readily understandable evaluations of effectiveness, providing policymakers and regulators with the basis for evidence-based decision making, (2) thanks to its meeting the interoperability requirement, fostering international cooperation, and (3) thanks to its addressing the challenges of defining and obtaining metrics for complex concepts and goals, advancing consensus around metrics and evaluation design.

3.2.3. Action to date

In recognition of both the benefits and the challenges of developing a benchmarking program that would meet the requirements we have identified, preliminary work has begun on the design and implementation of such a program. More specifically, under the auspices of the IEEE and The Future Society, a working group has been formed to explore the advisability and feasibility of pursuing such a project. The group includes representation from key agencies on both sides of the Atlantic. To date, the group has reached agreement on the need and the outlines of a program that would meet the need. Its current focus is on exploring practical questions related to how such a program should be developed. The group has not yet set a timetable for reporting on the results of its exploratory work.

4. Strengthening the AI operating environment through better tools

In the previous section, we considered a proposal aimed at strengthening the human environment in which AI is deployed through the fostering of a better-informed public. More specifcally, the proposal seeks to create a better-informed public through the establishment of an institutionalized program of open and interoperable AI benchmarking evaluations which have been designed to gather and publish sound evidence regarding the capabilities and limitations of AI-enabled systems when applied in real-world circumstances.

The evidence generated by benchmarking evaluations is a key input to a sound assessment of the trustworthiness of a technology. A well-designed benchmark (one accurately modeling real-world conditions, using data sets representative of those likely to be encountered in the actual application of a technology, and quantifying the various aspects of effectiveness through meaningful metrics) can tell us what we can reasonably expect (in terms of both capabilities and limitations) from a given technology in a given circumstance. That expectation can then be used to decide whether we have a plausible basis for trusting the technology to perform the task we are asking of it. The evidence generated by a benchmarking evaluation cannot, however, tell us whether the technology in question, once it has been applied, has in fact met its objectives in the specific circumstance in which we have applied it. If we want that information, we need to turn to *local validation*.

The results generated by a local validation exercise (a real-time or after-the-fact test of the effectiveness achieved by a given technology in a specific circumstance) are complementary to those generated by benchmarking evaluations. The latter tell us whether we have empirical grounds for believing that a technology of a given class will be successful in circumstances broadly similar to those modeled in the benchmark; the former tell us whether we have empirical grounds for believing that a specific instance of a technological system was successful in the specific circumstances in which we did apply it (specific data, specific hardware conditions, specific operators, specific timetables, and so on). Both questions are relevant in assessing the trustworthiness of a technology. The general question (answered by benchmarking evaluations) is most relevant before application, when we are deciding whether to adopt the technology for a given task. The specific question (answered by local validation) is most relevant after (or during) application, when we are deciding whether to trust the results that have actually been generated by the technology. Having reliable answers to both questions is essential to putting the adoption and use of advanced technologies in the service of the law on an empirically sound footing.

The complementary relationship between the two types of inquiry can be illustrated with an example taken from legal discovery in the US. The evaluations conducted in the TREC Legal Track (2006-2011)[35] produced results that showed that advanced retrieval technologies (often termed "technology-assisted review" or "TAR") *could* be reasonably effective at performing the task of retrieving documents responsive to a request for production.[6] That evidence gave responding parties the empirical basis they needed to adopt some variety of that class of technologies as the means to meet their discovery obligations (and, importantly, gave courts the empirical basis they needed to license that adoption). That evidence did not, however, obviate the need for local validation of the results generated by a given technology in a given matter. Requesting parties, and courts, still expect the circumstance-specific, after-the-fact, results that come only from local validation (and these expectations are often encoded in ESI ("electronically stored information") protocols which govern discovery procedures in a given matter). The general (TREC) evaluations provided the plausibility that gave the green light for adoption, but the matter-specific (local) evaluations are still needed to provide the evidence that establishes the soundness of the actual results.

4.1. The need

If local validation is an important element in an assessment of the trustworthiness of a technology, then there is a need to bring about the conditions needed to ensure that sound local validation exercises can be conducted often and everywhere. Here, however, there is a challenge. Whereas, in the case of benchmarking evaluations, the competencies required to design and run meaningful and statistically sound tests of the effectiveness of a technology can reside in a relatively small number of individuals (the individuals organizing and running the benchmarking program), in the case of local validations, those same competencies must be much more widely distributed. Individuals at a geographically very broad range of sites of AI deployment will need to be supplied with the competencies required to run meaningful tests of the technology as it has been deployed at their sites and in their specific circumstances. Meeting this need does not mean that every member of the public has to be equipped with the competencies required to design and run evaluations; it will suffice to widen the circle of competence to a broader range of domain experts. This is still a challenge however: how do we bring about a distribution of the required competencies that is sufficiently broad to answer the need for local validation?[7]

4.2. A proposal for meeting the need

If the objective of fostering a user pool better-equipped to gather evidence of the effectiveness of AI-enabled systems at the site of deployment is a worthy one, and if achieving that objective means bringing about a wider distribution of the competencies required to design and run sound local validation exercises, what might a solution that enabled that distribution look like? What we propose, and what we discuss in the remainder of this section, is the creation of a repository of resources that can be accessed by operators seeking guidance on how to design and run local validation exercises.

4.2.1. Requirements

If we wish to provide domain experts and operators with the resources needed to conduct local testing of the systems they are overseeing, the resources we make available to them must meet a number of requirements. Chief among these are the following.

- **Application-specific.** The testing that is required will vary from application to application. What is required for the local validation of an instance of TAR applied to the task of discovery, for example, will differ from that which is required for the local validation of a risk-assessment technology applied to custody decisions. The resources must therefore be application-specific and the ultimate goal should be the creation of a "library" of resources, each of which is tailored (in terms of test design, metrics, sampling procedures, interpretive guidance, and so on) to a specific task to which an AI-enabled system may be applied.

[6] The choice of modal is important here: the studies showed that TAR *could* achieve reasonably high levels of recall and precision; they did not show that TAR *would*, in all its instantiations and in all circumstances, achieve those results. Hence the need for local validation. This point is also sometimes insufficiently appreciated by readers of [39], which analyzed that showed that TAR *can* be superior to manual review (not that it *will* be superior in all circumstances).

[7] Of course, there will not be a need for local validation for every deployment of AI, but even restricting to deployments of sensitive applications, and even allowing for some level of aggregate testing of deployed technologies, there will still be a need for achieving a much wider distribution of the required competencies than we have today.

- **Tutorial and procedural content.** The resources should provide not only a procedural "recipe" for conducting a test, but should also provide sufficient tutorial content to enable an operator to understand the motivation behind a given procedural step (what a given term-of-art means, why a given metric is being used, why a given sampling design is chosen, and so on). To be effective, these resources should be calibrated for users with intermediate levels of expertise in the use and testing of advanced legal technologies. They need not be at the level of academic research papers, but they do have to go beyond elementary introductions.
- **Adaptable.** Even with the boundaries of a specific domain and task, there will be considerable circumstance-specific variation from one deployment of a system to another. The guidance provided by the resources should be of a sufficient depth to enable an operator to adapt the specified procedures for use in the specific circumstances at hand.
- **Intended audience.** The resources should be carefully calibrated to the level of expertise of their intended audience. Those who will be responsible for conducting local validation exercise will be a smaller, and technically more advanced, group than those consuming the results of those evaluations. The resources be calibrated to meet the requirements of these more expert users (while, to the extent possible remaining within the grasp, at least at a high level, of non-expert users).
- **Direction to other resources.** As a practical matter, the resources cannot cover every circumstance likely to be encountered in the real-world. While they should be of sufficient depth to cover the most common circumstances, they should provide direction to additional resources (including human resources) to consult when less typical circumstances are encountered.

What we have listed above are general requirements that any resource must meet if it is to serve the purpose of distributing the competencies needed to enable more frequent and effective local validation of AI-enabled systems. What we have not specified, however is any particular format for the resources. That is by design. There are, in fact, a range of different formats such resources might take (written procedures, glossaries, handbooks, video tutorials online calculators, and so on), and which format will be most effective will vary from one domain (and audience) to the next. We therefore leave the specific format as a question to be decided at the implementation stage.

4.2.2. Benefits

The creation of a repository of resources like that proposed in this section is no small undertaking; realizing the vision will require input from experts from a wide range of disciplines and subject-matter areas. The benefits of such a repository, however, would be considerable. These include:

- Improved competence;
- Improved effectiveness;
- Strengthened trust;
- Improved risk containment;
- More broadly distributed agency.

4.2.3. Action to date

The repository we have proposed remains, at the moment, aspirational; there is as yet no program under way to create it. Work has begun, however, on creating materials that would meet the requirements specified for resources in the repository and that could serve as a model for other resources.

More specifically, under the auspices of the IEEE and The Future Society, a project has been initiated, and in fact is nearing completion, to create a set of resources that, in the specific domain of legal discovery, will enable practitioners to conduct meaningful local validation of the results of applying advanced review technologies (or, for that matter, to the results of applying *any* review technology) to the task of legal discovery. The specific materials we have drafted are the following.

- **A Model Protocol.** An adaptable model ESI protocol that addresses the key issues that currently trouble parties in the discovery phase of litigation. The Protocol focuses on gathering the evidence needed to have an informed trust in the results of a review; its provisions are shaped by the principles of proportionality and evidence-based decision-making.
- **A Commentary.** A line-by-line commentary on the Protocol. The Commentary is designed to provide justification, interpretive guidance, and tutorial background for the Protocol's provisions.
- **A Handbook for Practitioners.** A companion document that provides an expanded discussion of the sampling and measurement procedures specified in the Protocol. The Handbook is intended to serve as a resource for advanced practitioners (and other stakeholders) seeking a deeper and more detailed understanding of the required statistical procedures.

These materials have been drafted and are currently being reviewed by a group of experts with a range of different perspectives on the use of advanced technologies for legal discovery and on how to put that use on the basis of an informed trust. We plan to publish the materials in 2023. Our hope is that the materials will both serve their immediate purpose of putting the use and testing of e-discovery technologies on a sounder footing and serve the larger purpose of serving as a model for resources that will enable the wider distribution of the competencies needed to conduct local validation of AI-enabled systems in other domains.

5. Concluding remarks

In this paper, we have drawn attention to the environment in which AI-enabled systems are deployed as a key element in any strategy for containing the risks (and for realizing the potential) attendant on the use of such systems. We have focused, more specifically, on the human component of the environment and considered two approaches (generating better information about AI's real capabilities and limitations, creating tools that will enable practioners to conduct local validation of the results of AI-enabled applications) for strengthening that component against risk. There are, of course, other aspects of the environment in which AI-enabled systems are deployed (hardware, software, even legal and financial) and exploration of ways to strengthen those components (making then more conducive to the detection, reporting, and resolution of risks) could pay off in more effective or efficient approaches to risk containment. One practical example is creating readily accessible pathways and repositories that would allow users (especially better-informed and better-equipped users) to report anomalies they have observed and to compare their observations with those submitted by others[4].

As can be seen by reviewing the requirements specified for the two proposals we have considered, the work required to strengthen the environment against AI-associated risk is non-trivial. To be successful, approaches on the environment-focused track require a considerable amount of planning, coordination, and effort. The benefits these approaches bring, however, are significant. Evironment-focused approaches may:

- By distributing more broadly the means for identifying and responding to unwanted outcomes from AI-enabled applications, avoid some of the adverse effects on innovation and technological development that may be occasioned by top-down approaches;
- By allowing practitioners to tailor solutions to their particular objectives and conditions, enable more nuanced and domain-specific approaches to risk containment; and
- By distributing knowledge more broadly (whether that distribution is direct or mediated by other entities or initiatives) advance the empowerment of the individual (against both private and state actors).

Given these benefits, we think that policymakers, and other stakeholders engaged in advancing the responsible use of AI, should always maintain an environment-focused (or bottom-up) track as a complement to the application-focused (or top-down) track. In fact, given the more benign collateral implications of environment-focused approaches, they should often be viewed as the solution of first recourse.

Acknowledgments

We would like to thank the organizers of 3rd International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023) for providing a forum at which we could express our views and hear those of others interested in these topics.

References

[1] M. Shanahan, Talking about large language models, arXiv preprint arXiv:2212.03551 (2022).

[2] E. M. Bender, A. Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5185–5198. URL: https://aclanthology.org/2020.acl-main.463. doi:10.18653/v1/2020.acl-main.463.

[3] OECD, Principles on AI, 2019. URL: https://oecd.ai/en/ai-principles.

[4] IEEE, Ethically Aligned Design, Version 1, 2019. URL: https://standards.ieee.org/industry-connections/ec/ead1e-infographic/.

[5] Council of Europe, European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment, 2018. URL: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.

[6] European Commission, Ethics Guidelines for Trustworthy AI, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[7] Future of Life, Asilomar AI Principles, 2017. URL: https://futureoflife.org/2017/08/11/ai-principles/.

[8] Partnership on AI, PAI Tenets, 2016. URL: https://partnershiponai.org/about/#tenets.

[9] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, M. Srikumar, Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai, Berkman Klein Center Research Publication (2020).

[10] T. Hagendorff, The ethics of AI ethics: An evaluation of guidelines, Minds and machines 30 (2020) 99–120.

[11] Y. Zeng, E. Lu, C. Huangfu, Linking artificial intelligence principles, 2018. arXiv:1812.04814.

[12] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al., Taxonomy of risks posed by language models, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 214–229.

[13] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al., Constitutional ai: Harmlessness from ai feedback, arXiv preprint arXiv:2212.08073 (2022).

[14] J. Mökander, J. Schuett, H. R. Kirk, L. Floridi, Auditing large language models: a three-layered approach, AI and Ethics (2023) 1–31.

[15] D. Long, B. Magerko, What is ai literacy? competencies and design considerations, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–16.

[16] P. Lin, J. Van Brummelen, Engaging teachers to co-design integrated ai curriculum for k-12 classrooms, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–12.

[17] D. Gašević, G. Siemens, S. Sadiq, Empowering learners for the age of artificial intelligence, Computers and Education: Artificial Intelligence (2023) 100130.

[18] Lepercq, V, Introducing Education, 2022. URL: https://huggingface.co/blog/education.

[19] Hugging Face, Introducing The World's Largest Open Multilingual Language Model: BLOOM, 2023. URL: https://bigscience.huggingface.co/blog/bloom.

[20] The Athens Roundtable, The Athens Roundtable: Artificial Intelligence and the Rule of Law, 2019. URL: https://www.aiathens.org/dialogue/first-edition.

[21] The Future Society, MOOC on AI and the Rule of Law, 2022. URL: https://thefuturesociety.org/2022/05/12/mooc-on-ai-and-the-rule-of-law\-successful-completion-of-the-pilot-phase/.

[22] aiEDU, aiEDU: The AI Education Project, 2023. URL: https://www.aiedu.org.

[23] C. McLeod, E. N. Zalta, Trust in stanford encyclopedia of philosophy, Metaphysics Research Lab, Stanford University (2006).

[24] IJCAI, Artificial Intelligence, 2023. URL: https://www.sciencedirect.com/journal/artificial-intelligence.

[25] AAAI, Association for the Advancement of Artificial Intelligence, 2023. URL: https://www.aaai.org/.

[26] IAAIL, International Conference on Artificial Intelligence and Law (ICAIL), 2023. URL: http://www.iaail.org/.

[27] NIST, Text REtrieval Conference (TREC), 2023. URL: https://trec.nist.gov/.

[28] Ministere de la Justice, Communique du Ministere de la Justice et de la premiere presidence de la cour d'appel de Rennes, 2017.

[29] World Economoic Forum, Responsible Limits on Facial Recognition; Use Case: Flow Management; Part II: Pilot phase: Self-assessment, the audit management system and certification, 2020. URL: https://www3.weforum.org/docs/WEF_Responsible_Limits_on_Facial_Recognition_2020.pdf.

[30] Snow, J, Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots, 2018. URL: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

[31] C. Garvie, A. Bedoya, J. Frankle, The perpetual line-up, Georgetown Law Center on Privacy & Technology 18 (2016).

[32] NIST, Face Recognition Vendor Test (FRVT), 2023. URL: https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.

[33] Center for Research on Foundation Models, Holistic Evaluation of Language Models (HELM), 2023. URL: https://crfm.stanford.edu/helm/latest/.

[34] NIST, AI Measurement and Evaluation - Panel on Measuring Concepts that are Complex, Contextual, and Abstract, 2021. URL: https://www.nist.gov/news-events/events/2021/06/ai-measurement-and-evaluation-workshop.

[35] TREC, TREC Legal Track, 2011. URL: https://trec-legal.umiacs.umd.edu/.

[36] LNE, METRICS - An international competition for the evaluation of robotics and AI, 2023. URL: https://metricsproject.eu.

[37] NIST, AI Measurement and Evaluation, 2021. URL: https://www.nist.gov/news-events/events/2021/06/ai-measurement-and-evaluation-workshop.

[38] S. Hallensleben, C. Hustedt, From principles to practice: An interdisciplinary framework to operationalise AI ethics, Bertelsmann Stiftung, 2020.

[39] M. R. Grossman, G. V. Cormack, Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review, Rich. JL & Tech. 17 (2010) 1.