INSTITUT FÜR INFORMATIK
der Ludwig-Maximilians-Universität München

TOWARDS MORE COLLABORATIVE AND ADAPTIVE ETHICAL REVIEW PLATFORMS

With the example of the ethical review of human computation-based citizen science projects

Libuše Hannah Vepřek
(Matrikelnummer: 10854333)

Masterarbeit

Betreuer Prof. Dr. François Bry
Abgabe am 20. September 2022

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst habe und keine anderen als die angegebenen Hilfsmittel verwendet habe.

München, den 20. September 2022 Libuše Hannah Vepřek

Abstract

Ethical review has become an integral part of research, especially in fields such as biomedicine and psychology. However, Institutional Review boards (IRBs) in the U.S. have, to a large extent, been tailored to fields that tend to focus on human experimentation, whereas fields like computer science have for a long time not been considered human subject research. Today, areas of research like Artificial Intelligence increasingly influence society, with related areas such as human computation-based citizen science raising entirely new ethical questions. So far, there do not exist sufficient established review and control mechanisms well-suited to these new ethical issues. Building on this need and on challenges presented by current IRB processes in general, this work discusses the question of how ethical review can be improved with respect to suitability, adaptivity, equity and efficiency. Focusing on the example of ethical review for human computation-based citizen science projects, it introduces, specifies and validates a new approach to ethical review in the form of CAER, a concrete design of a **C**ollaborative and **A**daptive **E**thical **R**eview system. CAER is presented with regards to the concrete personas and their needs that inform its design, a derivation of its core qualitative and technical characteristics, as well as a technical specification of key parts of the system. CAER goes beyond existing ethical review systems in allowing for dynamic and collaborative adaptation to the specific ethical considerations of different research fields, and improves the review process through semi-automatic evaluation of projects based on a rule engine and the adaptive ethical guidelines. This interplay of collaborative rule management by trustees and project representatives together with the semi-automated evaluation of submissions is itself a human computation system. By keeping humans in-the-loop during automated evaluation while at the same time continuously refining the underlying rule engine, the platform keeps adapting and improving. These two core design aspects (collaborative guideline management and automated review) of CAER are validated via user testing on proof-of-concept prototypes. In addition, insights from interviews with two IRB experts, regarding challenges around the IRB process and aspects of CAER's design, are presented. The results of the evaluation show that using CAER, the ethical review process can be more efficient than established procedures and that its adaptive design makes CAER potentially more suitable for various domains and research fields than existing systems. Furthermore, the testing of the prototypes suggests that no expert knowledge is required for using CAER, suggesting it may make the ethical review process more accessible while increasing its transparency. To fully explore the utility of CAER and to refine its new functionalities, further research and testing is required. Nevertheless, this work aims to lay the foundation for rethinking IRB processes and introducing a new collaborative and adaptive approach to ethical review.

Zusammenfassung

Die ethische Prüfung ist zu einem festen Bestandteil der Forschung geworden, insbesondere in Bereichen wie Biomedizin und Psychologie. Ethikkommissionen in den USA wurden jedoch weitgehend auf Bereiche zugeschnitten, die sich auf Experimente am Menschen konzentrieren, während Bereiche wie die Informatik lange Zeit nicht als Forschung am Menschen galten. Heute beeinflussen Forschungsbereiche wie die künstliche Intelligenz die Gesellschaft in zunehmendem Maße, wobei verwandte Bereiche wie auf Human Computation basierende Citizen Science völlig neue ethische Fragen aufwerfen. Bislang gibt es nicht genügend etablierte Prüf- und Kontrollmechanismen, die für diese neuen ethischen Fragen geeignet sind. Ausgehend von diesem Bedarf und den Herausforderungen, die die derzeitigen ethischen Prüfverfahren im Allgemeinen mit sich bringen, wird in dieser Arbeit die Frage erörtert, wie die ethische Prüfung im Hinblick auf Eignung, Anpassungsfähigkeit, Gerechtigkeit und Effizienz verbessert werden kann. Am Beispiel der ethischen Prüfung von Human Computation basierten Citizen Science Projekten wird ein neuer Ansatz für die ethische Prüfung in Form von CAER, einem konkreten Design eines kollaborativen und anpassungsfähigen ethischen Prüfsystem eingeführt, spezifiziert und validiert. CAER wird im Hinblick auf die konkreten Personas und ihre Bedürfnisse, die das Design bestimmen, seiner davon abgeleiteten qualitativen und technischen Kernmerkmale sowie der technischen Spezifikation der Schlüsselteile des Systems vorgestellt. CAER geht über bestehende ethische Prüfsysteme hinaus, indem es eine dynamische und kollaborative Anpassung an die spezifischen ethischen Notwendigkeiten verschiedener Forschungsbereiche ermöglicht und den Begutachtungsprozess durch eine semi-automatisierte Bewertung von Projekten auf der Grundlage einer *rule engine* und der adaptiven ethischen Richtlinien verbessert. Dieses Zusammenspiel von kollaborativem Regelmanagement durch Prüfer:innen und Projektvertreter:innen zusammen mit der semi-automatisierten Bewertung von Einreichungen ist selbst ein Human Computation System. Dadurch, dass der Mensch bei der automatisierten Bewertung im Spiel bleibt und gleichzeitig die zugrunde liegende *rule engine* kontinuierlich verfeinert wird, passt sich die Plattform ständig an und verbessert sich. Diese beiden zentralen Designaspekte (kollaboratives Richtlinienmanagement und automatisierte Prüfung) von CAER werden durch User Testing an Proof-of-Concept-Prototypen validiert. Darüber hinaus werden Erkenntnisse aus Interviews mit zwei Expert:innen zu Herausforderungen im Zusammenhang mit dem ethischen Prüfverfahren und Aspekten des CAER-Designs präsentiert. Die Ergebnisse der Evaluierung zeigen, dass der ethische Prüfprozess mit CAER effizienter sein kann als mit etablierten Verfahren und dass CAER aufgrund seines adaptiven Designs potenziell besser für verschiedene Bereiche und Forschungsfelder geeignet ist als bestehende Systeme. Darüber hinaus legen die Tests der Prototypen nahe, dass für die Nutzung von CAER kein Expert:innenwissen erforderlich ist, was darauf hindeutet, dass CAER die ethische Prüfung zugänglicher machen und ihre Transparenz erhöhen kann. Um den Nutzen von CAER

vollständig zu erforschen und seine neuen Funktionen zu verfeinern, sind weitere Forschungen und Tests erforderlich. Nichtsdestotrotz zielt diese Arbeit darauf ab, die Grundlage für ein Umdenken der ethischen Prüfverfahren und die Einführung eines neuen kollaborativen und adaptiven Ansatzes für die ethische Prüfung zu schaffen.

Acknowledgments

I would like to thank Prof. Dr. François Bry for the great opportunity to pursue the subject of this thesis as well as for his patience and guidance throughout the research and writing process. I am also grateful for his invaluable support throughout my computer science master's study.

My sincere thanks also go to Dr. Pietro Michelucci for introducing me to the interesting field of ethical review as well as for the enlightening and thought-provoking conversations and discussions and his tremendous support.

Moreover, I would like to recognize the invaluable assistance of my interview partners and test users and to thank them for their time and the incredibly valuable insights. Without their help, this thesis would not have been possible.

Finally, I would like to express my true gratitude to my family and particularly to my partner Philipp for always supporting and believing in me and encouraging me at all times.

Contents

CHAPTER 1

Introduction

1.1 Motivation and problem statement

For research to be ethical, scientific integrity and responsibility form the basic principles of good research practice. It is the duty of every individual researcher to commit to these principles and to conform to them in their everyday scientific practice [1]. Additionally, ethical review of research projects has become an integral part of research at universities, especially in fields such as bio-medicine and psychology to ensure that research projects and experiments do not harm human subjects and adhere to ethical guidelines. In the U.S., ethical review boards, or Institutional Review boards (IRBs), have been established for this purpose. Their ethical standards originated in the Nuremberg Code [2, p. 1331] and the Belmont Report from 1974 [3], as a reaction to severe ethical violations in research. These review boards have to a large extent been specifically tailored to fields that tend to focus on human experimentation [4, p. 115]. In contrast, fields like computer science and statistics had previously not been considered human subjects research [4, p. 115], but with the increasing influence of Artificial Intelligence (AI) on society and greater scrutiny from the social sciences, this is no longer sustainable.

This development has become particularly relevant to the branch of AI-research known as Human Computation (HC), which involves human participants as computational elements of larger systems. HC sets out to address problems which can not be solved by either humans or computers alone. The approach of HC has been especially successful, among others, in the field of citizen science (CS), for instance in projects such as *Stall Catchers*, where participants analyze Alzheimer's disease research data, or *Foldit*, where humans play a game to design proteins. This involvement of human participants is the reason that these CS projects require IRB approval, even though it is not always obvious whether these participants are in fact human subjects (as opposed to, for example, being researchers themselves [5], [6]). This new form of doing science also challenges current ethical standards of IRB review, such as the return of results, data openness versus data protection [7], data ownership and intellectual property [8].

Currently, there do not exist sufficient established review and control mechanisms customized to these new ethical questions and issues resulting from human involvement in

distributed thinking systems. Instead, HC-based CS projects are evaluated by the same boards responsible for biomedical and psychological research studies. This leads to certain problems: First, the ethical standards of existing IRB processes do not suit the specific needs of HC-based CS projects. Second, and this problem applies to the IRB process in general, accepted moral values and ethical standards generally change over time and given the advancement in AI research continuously introducing new forms of hybrid systems, new ethical questions may always arise. The ethical standards of today's classical IRB processes, however, haven't been updated since their introduction and therefore may often "not align well with emerging practices and technologies" [9, p. 165]. Third, human subjects—in this case citizen scientists—generally have no direct representation on the review boards themselves. Given that it is not clear yet which ethical issues are most important to consider in the field of HC and CS to protect the human participants, this is not sustainable. Lastly, and this—as will be described in more detail in chapter 4—also applies to the current IRB process in general, IRB review processes lack efficiency. Completed reviews are usually stored away and cannot be reused. If new projects with the same ethical issues apply for IRB approval, human IRB reviewers have to reevaluate all questionable cases from scratch, which is not only time-intensive but which also makes it even more difficult to establish consistent evaluations for new fields like HC.

1.1.1 Goals of this research

To contribute to the solution of these problems, this thesis introduces and discusses a new approach to ethical review with the example of HC-based CS projects, though the approach could easily be expanded and applied to many established and emerging research fields. The envisioned approach is presented as a concrete conceptual design of a collaborative and adaptive ethical review system (CAER System).

To be more precise, starting from the outlined problems with classical IRB reviews, the research question guiding this work is: How can we build a CAER System improving ethical review of HC-based CS projects with respect to the following aspects:

- 1) **Suitability** (regarding the specific needs and requirements of the reviewed project's research fields),
- 2) **Adaptivity** (regarding changing needs of the projects' stakeholders and knowledge gained from experiences in the fields over time as well as to respond to emerging technologies),
- 3) **Equity** (regarding representation and engagement of all stakeholder groups), and
- 4) **Efficiency** (regarding the order of review questions according to their impact on the review result and suitability for the project's field as well as the avoidance of redundancies).

The proposition underlying this work is that *HC itself* can help us realize such a system by leveraging the efficiency offered by computational methods while simultaneously relying on humans in the loop to ultimately control the evolution of the system and the ethical norms and standards represented within it. This is discussed in detail in section 1.3.

To answer the research question and to evaluate the suggested approach, this thesis is organized as follows. In the remaining sections of this introduction, the scope and relevance of the topic is discussed before providing essential definitions and some background information on the terms ethical review, CS, HC, and HC-based CS. Then, in section 1.3, an argument for why HC itself represents a valuable approach to enable collaborative and

adaptive ethical review is presented. In the next chapter 2, related work on review systems used for ethical review and research on improving ethical review in the U.S. is presented and existing software systems for ethical review and (collaborative) review systems as well as other related systems like Wikis and form-builder software, which provide helpful starting points, is discussed. To lay the ground for ethical review, the following chapters 3 and 4 then discuss current ethical review processes and their challenges. In chapter 5, the envisioned CAER platform is introduced, with a focus on the functional system requirements and technical specification. Chapter 6 presents an evaluation of the new CAER system based on both qualitative interviews with IRB experts and user testing of two proof-of-concepts prototypes implementing two of the core features of this new collaborative and adaptive ethical review process. Finally, chapter 7 discusses the broader (societal) impact of this new approach, its limitations and future work before the content of this thesis is summarized in the conclusion.

1.1.2 Scope and relevance of an ethical review process for human computation-based citizen science projects

This section frames the scope and relevance of ethical review forHC-based CS projects.

Despite the best intentions of researchers to not harm human subjects in their studies and to conduct ethically justifiable research, unintended impacts may always occur especially in new and emerging fields, where the needs of all stakeholders and norms and values may not yet have been fully established or sufficiently discussed. Ethical review helps to avoid such unintended consequences and ensures that research is conducted according to the current ethical guidelines, laws, and regulations.

Ethical review in the U.S. today is required for most research projects receiving funding for example by state funding agencies or for publishing results in specific journals. Although this is currently not the case in all countries, a trend in this direction can be observed that also makes ethical review of HC and CS projects also more relevant not only from an ethical but also from a formal perspective.

Finally, ethical review builds on a deep understanding of the field, the needs, norms, and values of the stakeholders, especially of the human subjects involved in a certain research project. For the field of HC-based CS, however, there do not exist established ethical guidelines like in biomedical research.[1] Since human HC-based CS also raises new ethical issues like the return of results or questions of exploitation [8], existing ethical guidelines do not fit the specific needs of this field. Nevertheless, these guidelines tailored to (for example) biomedical research are currently being used to evaluate HC-based CS projects. Thereby, ethical issues might remain unaddressed in the evaluation because of a lack of knowledge about the specifics of the type of research being conducted, while other issues may unnecessarily take up valuable time during evaluation, despite not being relevant to the research at hand. These challenges with the current ethical review system and IRBs in the U.S. are discussed in detail in chapter 4.

The establishment of ethical guidelines for the field of HC-based CS, as well as the development of a process to allow for the continued re-evaluation and evolution of such guidelines, are therefore essential for ethical review in general. More specifically within the scope of this work, these undertakings are vital to ensuring that research studies inviting

[1] Researchers such as David Resnik et al. [8] have therefore called for the development of guidelines for CS.

citizen scientists to participate in-the-loop in HC systems will not harm the subjects or participants involved and will align with the norms and values of the stakeholders involved.

1.2 Definitions and background

To provide the basic concepts for this work, this section introduces and defines ethical review, CS, HC, and, representing the intersection of the latter two, HC-based CS. Each term is introduced with a concise definition followed by more detailed background information which are necessary to understand the specifics of ethical review with the example case of HC-based CS.

1.2.1 Ethical review

Ethical review describes the evaluation and oversight of research studies by ethical review boards, also called independent ethics committee, research ethics board, or Institutional Review Boards (IRB) in the U.S. Ethical review ensures that research is conducted according to established ethical guidelines, regulations, and laws in order to protect the rights and well-being of human subjects as well as the integrity of the research process (see for example [10], [11], [12]). Most ethical reviews focus on research with human participants (see for example [13]). In contrast to, for instance, peer review, ethical review can, but in general does not, include scientific review. This means that ethical review boards can question the design and methodology of a study but usually only do so in the vein of ethics review [12, p. 4].

"Morality" and "ethics" commonly refer to the distinction between "good" and "bad" or between "right" and "wrong". In everyday life, norms of conduct guide and differentiate appropriate from inappropriate behavior [14]. Morals and ethics are sometimes being used synonymously but for the purpose of this work and following previous work [15] that builds on the understanding of the German philosopher Dietmar Hübner [16] of Ethics as a branch of philosophy, "Ethics" is seen as reflection of morality. Morality then consequently refers to systems of norms that may differ depending on their context such as different cultures or political currents [16]. These systems of norms have to be considered when deciding if a given moral abides to the common understanding of right or good behavior. Moreover, in most societies these norms exist besides legal rules and laws. This also allows to better understand the difference between ethical guidelines and morality and laws. The bioethicist David Resnik explains that "ethical norms tend to be broader and more informal than laws. Although most societies use laws to enforce widely accepted moral standards and ethical and legal rules use similar concepts, ethics and law are not the same. An action may be legal but unethical or illegal but ethical. We can also use ethical concepts and principles to criticize, evaluate, propose, or interpret laws." [14] The distinction between ethics and morals will be used to distinguish between norms and values in a given field such as HC-based CS and ethical guidelines that build on morality and are used to evaluate projects in the review process.

Besides education of researchers in the responsible conduct of research (RCR), which has become mandatory in almost all academic institutions in the United States [14], ethical review of research projects has become an integral part of many research fields at universities to ensure that research projects and experiments do not harm human subjects. These institutional or ethical review boards oftentimes have been tailored to specific fields with a focus on biomedical and psychological research and experimentation [4, p. 115].

Whereas ethical review of research projects is a global topic, the regulations vary from country to country, sometimes even locally. This thesis focuses on ethical review in the United States, although a certain degree of transferability to other national contexts is given.

1.2.2 Citizen science

"Citizen science" is generally understood as the "active engagement of the general public in scientific research tasks" [17, p. 1] with the aim of creating new knowledge [17, p. 22]. This minimal but widely accepted definition of CS excludes for example the participation in clinical trials or mere surveys in which participants do not play an active part in the study [18, p. 107]. However, besides this minimal understanding, there have been numerous attempts to define the term "citizen science". Different countries, but also individual organizations, hold their own definitions (for a list of different definitions see [19]). The analysis of definitions of CS conducted by [17] shows that most of these definitions remain vague and that the term itself is subject of debate. Other concepts that are often discussed under the umbrella of "citizen science" are for example "Open Science"[2] or "open innovation" [20, p. 2].

Although the term CS as it is understood today only emerged at the end of the 20th century, CS had already been practiced for centuries [20]. For example, we could refer to Charles Darwin as citizen scientist since he accompanied Captain Robert FitzRoy on the Beagle not as a professional naturalist but as an unpaid companion [21]. Another early example of CS before it received its own term and which is still practiced today is the Christmas Bird Count dating back to 1900 [22].

However, CS has especially seen a significant upswing and increase in participation over the last approximately 20 years, making it a "global movement" [20, p. vii]. This is due to, among other reasons, the spread of the internet which made it a lot easier for CS projects to reach a broad public, and for citizen scientists to learn about and engage in projects even from their computer at home. Over the last two and a half years, the growth of CS has further accelerated during the COVID-19 pandemic, which led to an increase in CS participation especially in projects dedicated to research relating to the coronavirus. For example, the CS project *Folding@home* experienced an increase of more than 970,000 active devices within less than three months in 2020 [23]

The engagement of citizens can take many different forms and vary from local to national or even international contexts and from short-term to long-term engagements [20]. For example, participants can either spend their own time to actively engage in CS projects or they can contribute in a passive way by providing their computing resources to projects such as *Rosetta@home* or *climateprediction.net*. The latter form of engagement is also known as volunteered computing [18] or voluntary distributed computing [24]. [18] has suggested a classification of participation and engagement in CS projects with four different levels: 1) "Crowdsourcing" following Jeff Howe [25] which includes volunteered computing and the engagement of "citizens as sensors" contributing data; 2) "Distributed Intelligence" in which participants contribute their cognitive ability to, for example, collect or analyze data, and carry out "a simple interpretation activity" [18, p. 117]. 3) "Participatory science" in which participants engage in the definition of the research problem and data collection method and also actively engage in the analysis with the assistance of institutionalized

[2]Here, CS is considered to be part of a larger "Open Science" framework of science. "Open Science" further includes approaches such open-access publications and accessibility of data [20, p. 8]

scientists. This form of participant is particularly significant in community science[3]. 4) "Extreme Citizen Science" which describes "completely integrated activit[ies]" [18, p. 117] in which participants together with professional scientists decide on research questions, methods, and the participants involvement. "Extreme Citizen Science" enables bottom-up approaches and creates the possibility for scientific projects entirely run by citizen scientists. [18] notes that these levels are not exclusive and some CS projects may include different levels of engagement for, for example, new and experienced citizen scientists. Overall, level 1 and 2 where professional scientists search for the help of citizen scientists to complete specific research tasks such as data collection or data analysis are particularly popular today.

CS is conceded to have the potential to not only "expand stakeholder participation and introduce new perspectives and information as well as new partnerships" [20, p. 2] but also to overcome the divide between science and society by opening up to the broader public to engage with but also learn about scientific projects [27]. At the same time, CS opens new perspectives for professional scientists and creates possibilities to communicate and share their research with the public and to increase its significance and impact on society [20]. Despite the great potential of CS, however, it also poses certain challenges, some of which have not previously been encountered in scientific knowledge production. For example, on the one hand and from the professional scientist's point of view, CS requires new ways of science communication (see different contributions in [20]) and requires new methods for quality insurance and validation. On the other hand, CS includes for example the risk of exploiting the volunteer participation by outsourcing usually paid tasks to volunteers without proper reimbursement.

In an attempt to provide a basis for "good" CS, [28] introduced the "10 Principles of Citizen Science" [28, p. 40]. These principles have since been widely accepted as "a common set of core principles" of CS across various disciplines and application fields. The principles emerged out of a working group of the European Citizen Science Association (ECSA) to underpin good practice in CS and the design of "responsible and impactful" [28, p. 28] projects. For the full list see [28, p. 29f.]. These principles—which refer to, among others, the mutual benefit of CS projects for both the professional scientists and citizen scientists, the scientific outcome of projects and to the different stages of projects, citizen scientists can participate in—form a good starting point for the development and evaluation of CS projects. However, these principles are guidelines on a rather high level of abstraction, leaving room for interpretation and consequently the need for ethical review systems that take into account the specifics and needs of CS stakeholders.

1.2.3 Human computation

HC describes an emerging research sub-field of AI which aims at "combining both the skills of humans and of machines [to] result in a higher problem solving competence both in quantity and quality." [29, p. 1]. Despite huge advances in AI in recent years there remain many tasks that humans can perform better than machines ranging from perceptual tasks such as visual image recognition to complex cognitive tasks such as heuristic problem solving [30, p. 2f.], [29, p. 1]. HC addresses precisely those problems that go beyond today's AI capabilities.

[3]Following [26] community science can be defined as "scientific research and monitoring, based on scientific modes of inquiry, which are (i) community-driven and community-controlled, (ii) characterized by place-based knowledge and social learning, collective action and empowerment, and (iii) with the normative aim to negotiate, improve and/or transform governance for stewardship and social-ecological sustainability." [26, p. 77]

The term "human computation" in its current understanding goes back to Luis von Ahn's doctoral thesis at Carnegie Mellon University in Pittsburgh, PA in 2005 [31]. In his thesis, von Ahn defined HC as "a paradigm for utilizing human processing power to solve problems that computers cannot yet solve." ([31] quoted in [32, p. 2]). HC is also known as "in-the-loop-computing" since humans take over computational tasks which include "the full spectrum of processes that might be applied to the transformation, synthesis, and interpretation of data" [33, p. 460] and can range from simple classification tasks to complex problem solving tasks such as protein structure design in the CS game *Foldit*. Hence, with HC, human cognition is being exploited within distributed information-processing networks [33, p. 459]. The contribution can be provided both consciously, as in the field of crowd-working or CS, or implicitly, as in the example of Google Maps [34, p. 2]. HC applications can, for example, take the form of so-called *games with a purpose* which "specifically target online gamers who generate useful data (e.g., image tags) while playing an enjoyable game" [30, p. xv] or identity verification mechanisms in which "users perform computation in order to gain access to some online content" [30, p. xv]. Bogner et al.[34] use the term "dual-purpose Human Computation systems" [34, p. 4] to describe HC systems which include two purposes: One purpose that motivates users to contribute and the "second purpose for which the systems take and process this information" [34, p. 4].

For an even more accurate understanding of HC, it is helpful to explain what it is *not*, and to distinguish HC from terms such as social computing and crowdsourcing, which oftentimes play important roles in HC applications and do have overlap with HC, but are not the same even if these terms are sometimes used synonymously in the broader discourse.

Quinn and Bederson [32] provide a thorough overview of HC and its differences to collective intelligence, social computing, and crowdsourcing. Here, I would like to only briefly mention those distinctions important for this work. In HC systems, users perform predefined computations to solve computational problems [30, p. 4]. HC can therefore include crowdsourcing, but is not synonymous with crowdsourcing which "replaces traditional human workers with members of the public" [32, p. 3] and not, as in the case of HC, "computers with humans" [32, p. 3]. For this reason, Law and von Ahn consider crowdsourcing as "a method or a tool that human computation systems can use to distribute tasks through an open call." [30, p. 4f.]. But, HC is not dependent on crowdsourcing and even not on a crowd or group of humans in general [30, p. 5] which is also why HC is not synonymous with collective intelligence [32, p. 4].

HC is also not synonymous with social computing, which is a widely defined term that encompasses a full range of social behavior and computer technology [30, p. 5]. Examples for social computing are blogs or wikis. The major difference between these systems and HC systems, according to Quinn and Bederson [32], is that "the purpose is not usually to perform a computation" [32, p. 3]. As in online discussions or "creative projects [...] the initiative and flow of activity are directed primarily by the participants' inspiration, as opposed to a predetermined plan designed to solve a computational problem." [32, p. 2]. The distinctions between these terms are not always uncontroversial[4] and remain open to interpretation especially with the development of new forms of HC systems which might question our current understandings of these terms.

[4]See for example [32] on the classification of *Wikipedia*.

1.2.4 Human computation-based citizen science

In the previous subsections, CS and HC have been treated as independent fields. However, CS has also been defined as a subsection of HC [29] or an application of it [33]. In this thesis, they have been treated individually first to provide a general overview of HC and CS and their broad application fields and to point to their different histories of origin. As should have become clear, CS approaches can also work without any computational components and hence without HC, and HC can also exist in other application domains that do not contribute to scientific research but follow other goals such as identity verification tasks. However, it is precisely the intersection of HC and CS where the combination of humans and computers has shown to be especially fruitful in terms of scientific knowledge production. For example, the HC-based CS project *Foldit* has already achieved successes in designing new proteins [35], hundreds of thousands galaxies have been classified with *Galaxy Zoo* [36], and *Stall Catchers* is advancing Alzheimer's disease research [37] [38]. One of the key strengths of HC-based CS according to Lintott and Reed [39] is that "[t]he willingness of large crowds of volunteers to give their time to projects that offer the premise of an authentic contribution to science, and the projects themselves are forming a critical piece of the response to the growing challenge of big data facing researchers in fields from astronomy to zoology." [39, p. 153] They offer solutions to the problem of handling vast amounts of research data that can be collected today but for which there do not exist computational solutions for their analysis. Moreover, HC-based CS allows for new creative solutions and/or serendipitous observations in the analysis of data, for example, in the case of *Foldit*, in the creation of new proteins which go beyond classical scientific approaches [39, p. 155]. Because of the described potential and proven successes of HC-based CS, the number and diversity of projects has increased significantly over the last 10 to 15 years [39, p. 161]. Since such projects are oftentimes developed at universities or are funded by state agencies, they fall under the scope of ethical review requirements. Here, they raise new questions regarding the roles of participants and challenge established ethical principles as is described in chapter 4. HC-based CS therefore presents a highly interesting example for the analysis of current ethical review processes in the U.S. and the development of a new collaborative and adaptive ethical review process.

1.3 Harnessing human computation to enable collaborative and adaptive ethical review

The proposition underlying this work is that HC itself enables a solution to the problem of ethical review of HC-based CS projects and beyond. Several reasons support this thesis and show how HC can contribute to solving the problem of ethical review of this (and other) field(s) to make the review process more collaborative, adaptive and suitable, equal, and efficient. Here, I would like to focus on four reasons in particular.
First, HC has proven to reduce computational time and to increase efficiency. For example, human contributions can strongly reduce the search space and hence the search time for computers [30, p. 11].
Second, HC can not only be more efficient than purely computation solutions, but also more effective. Michelucci [33], for instance, shows how HC increases effectiveness via coordinated action with the examples of WikiProject and Swarm theory. Wikiproject coordinates "the online activities of Wikipedia contributors by leveraging the availability of user activity information" [33]) and thereby decreases duplicate efforts. Based on swarm theory, Michelucci argues that "the opportunity for centralized coordination afforded by mediating technology could be used to share locally relevant information with individual actors to improve their efficacy, and also to implement more complex activities." [33, p.

14].
Third, ethical review is a collaborative effort and, as [40] states, IRBs "are more than the sum of their parts" [40, p. 22]. Different work has shown how a HC-based approach can be utilized for addressing collaborative challenges (see for example [41], [42]).
Finally, and this seems to be the most important reason for the development of a collaborative and adaptive system for ethical review, HC ensures human control as humans retain ultimate decision-making power. In comparison, purely computational AI systems that rely solely on machine learning for example—despite increased efforts towards explainable AI—most often lack transparency and interpretability of decisions reached and the underlying model. In HC systems, humans "drive the societally-relevant decisions and behaviors of the system." [7, p. 11]. Bowser et al. [7] aptly describe this advantage of HC: "The conscientious development of AI systems that carefully considers the coevolution of humans and technology in hybrid thinking systems will help ensure that humans remain ultimately in control, individual or collectively, as systems achieve superhuman capabilities." [7, p. 11]
In this work, HC is presented as a viable approach to ensure that humans remain in the loop and that different stakeholders can contribute on an equal basis. For these reasons, this work builds on HC to develop a new approach to ethical review.

CHAPTER 2

Related Work

This work can draw from research and many different existing software systems built to digitize ethical review. There exist various prior studies and attempts to improve ethical review, a brief overview of which is given in this chapter. Since the envisioned system at the center of this work has overlap with other applications, such as review systems in general, salient examples of these are also discussed in this chapter. At the end of this chapter, existing software systems that focus on specific aspects that could be helpful for the development of the new CAER platform are mentioned. These existing systems form a valuable starting point for this thesis, however, there currently exist no software systems and only a few theoretical research studies that seek to improve ethical review by rethinking its fundamental process in general. Attempts at introducing a HC-based approach to this problem, as suggested in this thesis, have also not been encountered in this research into related work.
In sections 2.2-2.5, various existing software systems that are of interest for this work including peer review, wiki software and form builder software are discussed. These sections also include notes on how various features of these systems could usefully inform the design of the envisioned CAER system. The details of some of these assessments may appear as not-yet fully justified, as some of the requirements and shortcomings for CAER that are alluded to are derived and explained in detail in other sections of this work, particularly 4 and 5. The choice to nevertheless discuss related work earlier on was made to provide a more convenient flow to the reader, as well as to make clear what capabilities may or may not already be provided by existing systems before finally specifying CAER.

2.1 Review systems used for ethical reviews

As previously discussed, ethical review processes differ around the world, but in accordance with the focus of this thesis, the main focus of this section is on ethical review systems as used in the U.S. First, existing software systems currently used by IRBs in the U.S. are discussed. In the second part, research on the improvement of review systems with computational solutions or efforts to build new review processes is presented. These approaches refer to limitations of current IRBs and the ethical review process in the U.S. described in chapter 4.

2.1.1 Existing software systems for ethical review processes

IRBs today commonly use computational systems to manage ethical review processes. There exist multiple such systems, of which several representative examples are presented in the following.

There exist different IRB software systems provided by private companies such as the *Cayuse* cloud platform [43], the IRB software *Axiom Mentor* [44], *OneAegis* (formerly known as *IRBManager*) by Tech Software [45], *InfoReady* [46], *Huron IRB*, and *Key Solutions software* [47]. These platforms mostly comprise the same features with the aim to cover and improve the whole IRB cycle.

For example, *OneAegis* (formerly known as *IRBManager*) by Tech Software is an electronic system for IRB application submission and management used by various IRBs in the U.S. which was originally developed in collaboration with the Biomedical Research Alliance of New York (BRANY) and assists IRBs with a management and notification system for the review process [45].

Another widely used example is *Huron IRB* which also aims at facilitating the IRB process by providing management and collaboration features and with *Huron IRB Exchange* additionally allows to connect Single IRBs form multi-site research [48].
InfoReady [46] is a platform to streamline various review processes such as research administration, scholar- and fellowships, journal and conference review processes and has lately been applied as software for the IRB process by the Eastern Kentucky University [49]. The advantages of *InfoReady* focus on improving the review process by facilitating online submission and access to all resources, providing a history a review history in one place and automated assignment of tasks [49].

Key Solutions software [47] offers another solution for facilitating the IRB process, with its key features being centralized access, reporting, integration with other modules, and automation of the approval process. While this suggests a fully automated approval process, it actually refers to "routine tasks such as meeting management, agenda setting, document routing, and e-mail notifications [that] are automated by the product. An automatic 'Check for Completeness' is performed before protocol submission to ensure that the application is complete." [47]

Regarding the use of smart forms, for example the IRB software systems *Cayuse* and *Axiom Mentor* [44]) include smart forms in the IRB cycle to create forms specific to a certain research field. However, the smart form functionality of *Axiom Mentor* focuses on mentor-usage. Similar to *Cayuse*, mentors have to create these custom tools themselves, for example by using drag and drop to create new forms. *Axiom Mentor*, just like some of the other software systems, comes with the advantage that clients do not have to worry about physical infrastructures, which is all cared for by the platform provider themselves.[1]

Furthermore, besides these generic IRB software systems, there also exist software systems that are specifically tailored to clinical research and include IRB services, such as *ADVARRA* [50]/) or the web-based e-clinical solution *Trial Interactive* [51]. However, since this research does not focus on clinical research and the existing software systems do not appear to address the question of how to adapt ethical review to fields of research other than biomedical or clinical studies, for example, these systems form interesting reference points for features such as communication and notification features, but do not assist to an-

[1] For a detailed list of *Axiom Meter*'s features see [44].

swer the question of how to make ethical review more suitable for emergent research fields.

It should also be noted that there exists a large number of different custom IRB software systems, as many institutions have developed their own IRB systems, such as *eIRB+*, the Northwestern University's "electronic submission and review system" [52], to give one example. Such systems were not found to have been described as including any automated review functionalities as developed in this work. Another example is *ERICA*, the University of Utah's Electronic Research Integrity & Compliance Administration system [53]. *ERICA* also includes smart forms that provide application pages according to the study to be reviewed but does not connect this with any form of automated evaluation.

Finally, other interesting efforts focus on the development of tools that facilitate multisite research such as the *Streamlined, Multisite, Accelerated Resources for Trials IRB Reliance platform* (*SMART IRB*) [54], which is specifically developed to integrate and facilitate the IRB process for multisite studies and as such is not an IRB review system but a master reliance agreement [55]. Other examples are the *Adverse Event Data Management System* (*AEDAMS*) [56], which focuses on managing and monitoring adverse events but includes features to generate protocols for IRB submission, and *IREx* (*IRB Reliance Exchange*), which offers a "freely available, web-based portal supporting single IRB review documentation and coordination for multi-center clinical trials" [57].

Overall, the described software systems present great support for current IRB processes and streamline the workflow of state of the art "classical" ethical review, but do not address the challenges IRBs currently face (described in chapter 4). For example, they do not address the review of research in emergent fields which are not covered by current guidelines and regulations. They also do not allow users to collaboratively update any ethical guidelines accordingly. Even though some existing IRB software systems include smart forms, these have to be created and arranged by the IRB members in a manual way, effectively creating two separate "sources of truth", namely the ethical guidelines themselves, and the forms with which to evaluate projects against these. Finally, it can be argued that the field would benefit from open-source software systems instead of proprietary software, to increase transparency and build trust in the ethical review process.
In the following section, existing research on and efforts to improve current systems and the ethical review process in general are outlined.

2.1.2 Research on the improvement of ethical review processes

In his recent article on "Institutional Review Board Decision Support Software" Stephen Rosenfeld states that

> "[t]o date, most software used to support IRBs has focused on the problem of collecting submission documents from sponsors and investigators. Stacks of protocols, consents, and investigator brochures have been replaced by workflow and document management systems that have greatly reduced the administrative burden on IRB members and staff and improved review process performance. Little concern or attention has been focused on what IRB members do with these documents beyond providing regulatory checklists to promote and document compliance. This lack of attention to tools that can improve decision-making is a missed opportunity to manage quality." [58, p. 10]

Even though the problem of availability of materials has been solved with existing IRB software, there still remain many problems and room for improvement for such computational systems [58, p. 11]. The author focuses on the three areas "context, history, and

precedent"[58, p. 11] to suggest how IRB software could be improved. First, to provide more context for decision-making, tools could include search options and links to material on websites such as *clinicaltrials.gov* which are relevant for the ethical evaluation of research [58, p. 11]. Second, software could provide "a bird's-eye view of prior reviews" [58, p. 11] to make decision-making more reliable and consistent by allowing IRBs to easily access a research history and progress. Finally, access to previous decisions could be facilitated as well as options to compare different decisions to increase consistency of decisions on related issues [58, p. 11]. These suggestions remain theoretical, but Rosenfeld has recently applied for funding to develop a cloud-based IRB software tool called *Research Metadata (RM)* that builds on these design suggestions.[2]

To make the review process more efficient and faster for minimal-risk research, [59] conduct research on developing a web-based decision tool, a so-called *wizard* to evaluate if a research study is exempt. The tool can also help to identify need for additional review [59]. After the first proof-of-concept [60], the *wizard* has been updated and revised to, amongst other things, include more review-critical details, update exclusionary criteria and to refine the definitions of terms such as "human subject" [59].

With *Bartleby*, [61] have developed a system that helps researchers conduct ethical research. The open-source software provides debriefing for large-scale online research and allows participants to opt out or to demand for their data to be deleted [62]. *Bartleby* especially supports discussion of the role of research participants and their voice and empowers them to express autonomy [61]. This work unquestionably describes important paths to more ethical research but does not address ethical review.

In 2015, a study by Karl Oder and Stephanie Pittman [63] analyzed the impact of IRB computer systems on the number of staff members needed and questioned if IRB computer systems would actually facilitate the process. The authors conclude that usage of a computers system does not have any effect on the number of IRB staff members, but that the number is more correlated with the number of protocols reviewed by an IRB office and that offices with more protocols tended to use a computer system. It has to be noted that this finding presents only a single study on the effect of computational systems on the IRB process. Although the study should not be omitted and the results kept in mind, the huge number of different software systems and attempts to improve them shows, that there is a lot of interest in computer systems to aid the IRB process and that they are experienced to be of value.

In addition to studies such as those already mentioned, the overall awareness of the need for ethical review of computer science research is growing. This can be seen for example in NeurIPS' (The Conference and Workshop on Neural Information Processing Systems)[64] efforts to establish an ethics review process for submissions to the conference and lately to develop a "NeurIPS Code of Ethics" [65] together with the NeurIPS community.

Finally, two initiatives to build new ethical review boards to improve the ethical review process should be mentioned.

First, Bernstein et al. [66] developed and evaluated the "Ethics and Society Review Board (ESR), an institutional process that facilitates researchers in mitigating the negative societal and ethical aspects of AI research" [66, p. 2]. The *ESR* focuses on societal harm

[2]Information received from email exchange with Dr. Stephen Rosenfeld on June 14, 2022.

of AI systems and builds on an academic board of researchers and experts from various field. However, the *ESR* can also be said to not sufficiently include the affected communities themselves.

Second, with the aim to create a new approach to IRB review called "Learning IRB" which builds on "continuous learning for both operational efficiency and decisional quality" [67], Stephen Rosenfeld together with Patricia Seymour and colleagues founded the *North Star Review Board*, a Research Ethics Review Board (RERB), in 2021. With this approach the review board wants to meet the consideration that science and the development of technology is not static but changes over time and that research on AI for example raises new questions that have not been answered so far. The RERB therefore specifically focus on open ethical questions [67]. The review board is deliberately nonprofit to build trust in their decisions as independent from institutions or investors. This approach is very promising and the new CAER system suggested in this work could further support this model by facilitating specific processes by ways of HC-based automation.

2.2 Review Systems

After having discussed research on and systems for ethical review in the U.S., the following briefly discusses new research in the area of peer review and technology-assisted review systems. It thereby particularly focuses on two recent publications.

With the goal of reviewing research projects with respect to their ethics, ethical review shares similarities with peer review. The goal of peer review in general is to assess research in terms of its "competence, significance, and originality." [68]. It aims to ensure quality control to reduce "misinformation and confusion" [69] thereby upholding the integrity of science and the public trust in science [70]. It also ensures the quality of the published research. In the presence of an overwhelming number of papers written, peer review also takes another role, which is to help select and filter research of qualitative value [71, p. 76]. However, peer review today also comes with challenges which have been discussed by [72]. The author considers current challenges of peer review, provides examples for experiments to better understand the problems and presents computational and formal or organizational solutions and their impacts found in literature. The focus of this review lies on peer review in scientific conferences although many aspects also apply to other forms of peer review such as journal review or review of grant proposals [72, p. 78]. In the common review process, each paper submitted by authors is reviewed by several, usually three to six reviewers, whereas the number of papers assigned to reviewers may vary. Reviewers evaluate the assigned papers and provide suggestions for improvement within a set deadline before the rebuttal phase, in which reviewers can update their reviews. This phase is followed by a discussion between reviewers and meta reviewer before the meta reviewer finally presents a recommendation for decision about the acceptance of a paper. The final decision lies with the program chairs [72, p. 78].

[72] identifies challenges (and possible solutions) in the outlined peer review process which include challenges in the reviewer-paper assignments. Present solutions to improve the matching of reviewers and their expertise to papers use for example natural language processing techniques [72, p. 79]. Moreover, issues exist also in the bidding process, in which reviewers can place positive, negative or neutral bids on papers to review, which can partly be prevented with algorithms that order the presentation of papers in the bidding process in a dynamic way [72, p. 79]. However, Shah also points to trade-offs of

solutions that for example assign papers to reviewers by maximizing the similarity score of assigned reviewer-paper pairs, which can also lead to unfair treatment of some papers [72, p. 80]. Another challenge identified by the author is dishonest behavior, for example by providing poor reviews to increase the chances of one's own authored paper. This problem, however, usually is not present in ethical review, since researchers with to-be-reviewed projects commonly do not engage in the same ethical review board and there exists no limit in the amount of research projects that may be accepted, i.e. pass ethical review. Additionally, miscalibration, or the fact that the meaning of ratings are subjective, forms another challenge [72, p. 82]. This problem can be partly circumvented, for example, by providing rankings of the reviewed papers [73], [74], [72, p. 83], but still remains a subject of debate. Finally, subjectivity in the review process and group dynamics in the discussion of papers form open challenges. For example, review criteria are often interpreted differently or weighted differently by reviewers [72, p. 83]. Bias regarding author identity can be addressed with single-blind or double-blind review, though each solution comes with its own challenges. [72] concludes that despite various attempts to improve peer review, this research, "particularly using computational methods, has only scratched the surface of this important application domain."[72, p. 87]

Finally, and particularly interesting for this thesis could be the "open review" approach which has recently been pursued by some conferences, where both the reviews and corresponding submitted papers are published while the reviewer identities remain anonymous. *OpenReview.net* is an example for a conference management system applying this approach [72, p. 85]. Such an open review approach improves transparency "and provides more information to the public about the perceived merits/demerits of a paper" ([72, p. 85]. However, the approach could also lead to bias in review if a reviewer knows that a paper is being resubmitted, but this issue might be less severe in ethical review, since re-submitting projects to another IRB is common practice for ethical review and IRBs are commonly aware of this.

Since ethical review shares similar challenges to peer review, considering prior attempts to improve peer review is important for informing this work. However, in practice, ethical review has not been part of peer review. In some fields, ethical review might have always played a role, but it has not been considered a core aspect of the review process in general. The latest developments at the NEURips conference discussed above show the growing awareness for including ethical review in peer review processes.

Other important work such as the workshop on "Augmented Intelligence in Technology-Assisted Review Systems (ALTARS 2022)", which took place in 2022, particularly discussed evaluation metrics and protocols [75] in the context of High Recall Information Retrieval Systems. This important work is very interesting for this thesis, since technology-assisted review systems make use of HC approaches "where classification and/or ranking algorithms are continuously trained according to the relevance feedback from expert reviewers, until a substantial number of the relevant documents are identified." [75, p. 558]. The workshop in 2022 was the first of its kind and points to the topicality of finding and incorporating novel approaches, (computational and processual) to review, and learnings from this line of research are likely to be applicable to ethical review in similar ways as to other kinds of review processes.

2.3 Collaborative review systems

While existing software systems for ethical review as presented above present very specific collaborative review systems, this section focuses on collaborative review systems in general, which include features worthwhile to consider to include in ethical review systems as well. There exist several such collaborative review systems with *Easychair* [76], *Microsoft Conference Management Tookit* (*CMI*) [77] and *Open Journal Systems* (*OJS*) [78] being some of the most common solutions. Other systems for collaborative paper peer reviewing and journal management are *Scholastica* [79], *Hindawi* [80], *OpenReview* [81], *OpenConf* [82], and *Indico* [83]. In the following the focus will lie on the three most common systems. While *Easychair*, *Microsoft CMT* and *OJS* are all review and publishing software, *Easychair* and *Microsoft CMT* are also conference management software.[3]

Easychair is a web-based conference management system with the main functionality being conference management and organization (including for example registration, publishing conference proceedings and preprints, and most important for this thesis, reviewing). Moreover, a journal reviewing and publishing platform is under development (August 2022) [76]. *EasyChair* therefore provides a solution to plan, organize, run a conference, and publish its proceedings in one environment. It also supports different models for conferences with a single program committee and for conferences with multiple tracks. Here, particularly those features that are relevant for a collaborative review system should be mentioned. *EasyChair* includes:

- Account and membership management
- Communication features (emails)
- Role management (changing roles possible (author, reviewer, sub-reviewer, Program committee member)
- Templates to create for instances a call for papers, to write a review etc.
- Submission of papers
- List of submissions; list of all projects and their status
- Review management includes Review request functionality (admin can request review by specific reviewer)
- Online discussions of papers
- Author-rebuttal phase which allows authors to respond to the reviews
- Conference statistics

The second examples is the *Conference Management Toolkit* (*CMT*) by Microsoft Research which includes a full submission life cycle from abstracts to full papers and presentation. It has been launched in 2016 and has hosted over 7000 conferences since then [77]. *CMT* runs on Microsoft's Azure cloud platform and is accessible through web-based interfaces. Amongst *CMT*'s features, the following are particularly of interest for the topic of this thesis:

- (Multiple) role and user management (besides reviewer also meta-reviewer etc.; assign tasks to individuals etc.)
- Review cycle with submission, review and presentation of papers

[3]The following descriptions mainly rely on the platform's documentation since it was not possible to obtain trial accesses for the purpose of this master thesis.

- Conflict management (helpful to ensure that no trustee affiliated with a project to be reviewed will receive the review)
- Communication features (email, automated notification system, etc)
- Review management: allows to set date for expiration of invitation to review (if not answered)
- Meta-Review
- Discussion feature
- Rebuttal feature (authors can react to review)
- Activity log
- Customizable forms

CMT also forms an interesting starting point for the system outlined in this thesis, with many similar features to *Easychair*. Here, those features that are different from the other systems described in this section should be pointed out: The (meta-)reviewing and discussion feature, which allows to initiate a discussion and ask specific reviewers to discuss a certain issue and to come to an agreement, are a helpful addition. Additionally, *CMT* includes conflict management functionality is useful in ensuring that disagreements between submitters and reviewers can be addressed inside the platform. Finally, *CMT* features customizable forms, allow users to create their own review forms.

Finally, *Open Journal Systems* (*OJS*) is an open source publishing platform for "managing and publishing scholarly journals"[78] (released under the open source GPL v2 license), which has been developed by the multi-university initiative Public Knowledge Project based at Stanford and Simon Fraser University. According to *OJS*' website "it is the most widely used open source journal publishing platform in existence, with over 25,000 journals using it worldwide." [78] Their goal is to make open access publishing a feasible option for different journals. *OJS* covers the whole workflow of publishing journals and includes the following features amongst others:

- Creating a journal website
- Flexible and configurable editorial workflow (for instance, multiple sections are also possible)
- Online submission, management and archive
- Tracking features to keep track of the work of authors, reviewers, editors
- Assign tasks and submissions to users (e.g. reviewers)
- Communication features which allow to contact individual users or to notify e.g. specific roles [78]
- "Pre-review discussions"; Functionalities that allows to have a discussion with everybody involved in a submission
- Dashboard including an "expanded task menu" for all tasks and their status

OJS is extensible and customizable due to its open-source approach. Moreover, it introduces a multiple sections feature that allows users to create different sections for a journal. Of special interest is the reviewing process here, which is not only very flexible but also includes communication functionalities between authors and editors. This allows authors to answer questions that come up in the review process but also to comment on the review

process themselves and to ask questions about it to further improve their work.

In conclusion, the three reviewed software systems for collaborative review describe very interesting starting points for this thesis and include many features that are important for developing the CAER platform. For example, the submission of papers that all systems feature could of course be seen to correspond to submission of research projects for ethical evaluation in CAER. Similarly, list views of submissions could represent all to-be-evaluated research projects and their status. Moreover, *Easychair* features human review subsequent to automated review, and a similar requirement can be derived for CAER. *CMT* also has a meta-review feature and conflict management, learning from both of these could usefully inform the design of CAER. Another example is the "multiple sections" feature in *OJS*: a similar approach to this could be adapted in CAER to cluster ethical issues topically (but this is not done as part of this work). Finally, common features such as user and role management, communication and statistics around submissions would also be beneficial in CAER and are featured by all of these systems in different ways.

However, these tools still have some shortcomings with regard to ethical review with the envisioned CAER system seeks to address, and which will be the primary focus of its specification in this work. Of course, this is to be expected, since these systems have been designed and built for different purposes than ethical review. One important aspect to keep in mind is that the modality of evaluation is fundamentally different for the review of scientific papers than for the ethical evaluation of research projects. This will become clear throughout this thesis (specifically, in chapter 3 on the ethical review process) and is also evident from the fact that various attempts to build standalone software systems for ethical review exist (see above).

Moreover, the goal of the new review system developed in this thesis is to improve the current review process by taking load of the shoulders of human reviewers by introducing an automated review system that initially "reviews" the project based on current ethical values. In contrast, with *EasyChair*, recurrent reasons for rejecting a paper have to be described for every single paper. In ethical review by contrast, if it is clear that a certain issue leads to the rejection of a paper or project, this could be included in the automated system. The reviewed systems do not include or allow the automation of particular steps of the review process. Nevertheless, with features such as communication and task assignments, the reviewed systems form a great inspiration and starting point for this research.

2.4 Wiki software - MediaWiki

In this subsection, *Wikipedia*, or rather its underlying software *MediaWiki* will be discussed. The reason for including Wiki software in this review of existing systems is that their collaborative approach to continuous modifications of content is relevant for a collaborative review system that builds on the understanding that moral values and ethics change over time.

Launched in 2001, *Wikipedia* is the world's biggest free, online content encyclopedia that "anyone can edit" [84]. It is supported by the Wikimedia Foundation "and based on a model of freely editable content" [84]. In general, a mostly anonymous group of contributors collaboratively create *Wikipedia*. This means that, technically, anybody can contribute, as long as they abide to *Wikipedia*'s policies and the content they contribute is free of any copyright restrictions.

What makes *MediaWiki*'s approach powerful and interesting for the envisioned review system is that the software enables quick and easy reversal of wrong content and mistakes so that it's continually being updated in a collaborative way. In the following, some of *MediaWiki*'s features will be pointed out that are particularly interesting for the envisioned collaborative and adaptive ethical review system:

- Articles are never complete but based on a "live collaboration", which means that they are constantly created and kept up-to-date [85]
- With the "Editorial Quality review" Wikipedia also includes specific control structures to ensure that questionable content cases, for example, can be resolved
- It includes a "page's history" which allows to track changes by individual contributors and is publicly available
- Cross-references of articles
- Discussion pages ("Talk pages") for individual articles or a "WikiProject"
- Communication features (in-platform: contact individual contributors via their talk page, IRC and e-mail, etc.)
- Contributors can form "WikiProjects", which are working groups that work together on certain topics
- Dashboard provides an overview over ongoing discussions or current requests

Even though *MediaWiki* has not been developed for ethical review, some of these features could provide useful inspiration for the envisioned ethical review system. The possibility to continuously create and update articles is very interesting, since the envisioned dynamic review system needs to make it possible to dynamically and continually adapt ethical issues and rules over time, and collaborative editing is one way to enable this. Similarly, "talk pages" enable the discussions on issues and rules and with "WikiProjects" different stakeholder groups could be formed.

However, since *MediaWiki*'s purpose is different from ethical review, it has some shortcomings when it comes to the specificities of the ethical evaluation of research projects and may not be suitable as an off-the-shelf system for ethical review. For example, Wiki software neither has a native notion of feature-based or categorical reviews, nor does it include any form-based logic/rule-application mechanism that could be used to implement reviews as such to begin with. Additionally, ethical review includes very sensible information and questions and edits made to the ethical guidelines should therefore only become visible if they have been approved by several trusted members and IRB experts, which may be difficult to properly implement with *MediaWiki*'s access control mechanisms. Wiki software nevertheless contributes important inspiration for the envisioned CAER system.

2.5 Form builder software

Form builder tools, especially those that allow the creation of so-called "smart forms", implement much of the functionality that is needed for development of the automated evaluation of projects envisioned in CAER[4]. "Smart forms" are forms which adapt dynamically according to previous input by a user.

[4] As noted earlier in this chapter, a more detailed derivation and justification of this and other requirements is presented in chapters 4 and 5.

There exist various online form builder tools that allow users to create and share customized (smart) forms without requiring any coding. Because of the large number of providers of such software with comparable functionalities, only four examples that were found to be quite representative are briefly discussed in the following. For more online form builders see for example [86], which introduces various online form builders with partly different focuses on for example security and encrypted submission.

The first example is *Airtable*, a low-code platform for building collaborative apps [87]. A distinctive aspect of *Airtable* is that it combines functionalities of relational database systems with those of spreadsheets. Here, it is of interest for the development of a new ethical review process since it allows to build and customize smart forms efficiently. Being web-based, *Airtable* enables sharing forms an collecting data easily via shareable URLs. Information from submitted forms can automatically be stored in a table (inside a so-called "Base"). Moreover, with conditional logic, form fields can made to be displayed only depending on values chosen by the submitter in previous fields [88]. This way, only relevant fields will be presented to the submitter, reducing the time needed to fill the form. The use of required fields ensures that submitters submit their form only after having filled in all important information. Note that this type of application of conditional logic to alter the flow of a form is essentially what is taken to be the definition of a "smart form" in this work. Smart forms are relevant in the context of CAER as they offer a way to potentially speed up the submission process by only requiring participants to answer questions that are contextually relevant for the evaluation of their projects.
Airtable presents one suitable solution for building such smart forms for the automated ethical evaluation of new HC-based CS projects. In fact, due to its relatively high versatility in implementing not just smart forms but also relational tables and basic user interfaces (UIs) to interact with them, it was chosen as a basis to implement the prototypes for the validation of some of CAER's key features presented in 6.

Google Forms is the second example for form-builder software that will be briefly discussed here. With *Google Forms* it is possible to generate surveys in the web browser and to analyze them instantly with the help of charts and graphs provided by Google [89]. One can create blank forms or use one of the templates. *Google Forms* includes the possibility to include conditional questions which only show up if another question has been answered and to set questions to "required" to answer. However, *Google Forms* does not support conditional logic in the same way that, for example, *Airtable* does. Importantly, it is not possible to specify multiple conditions connected with logical "and" and "or" operators.

Jotforms is another example of form-builder software, allows users, among other things, to generate reports, to capture payments, and to trigger automated workflows [90] based on conditional logic. For instance, it is possible to present different PDF downloads or to send messages to different addresses depending on the user's answers [91]. *Jotforms* also allows to combine several question/answer-pairs to one condition.

Last but not least, *Typeform* is a form-builder software that includes conditional logic and presents one question at a time to keep users engaged [92]. With its "Logic Map" tool it allows to easily modify and include conditional logic in a visual way. *Typeform* furthermore allows to group related questions together in question groups.

The research on and existing systems for ethical and collaborative review, Wiki software, and "smart" forms described in this chapter form important foundation for this thesis. As will become clear throughout this work, a new dynamic and collaborative ethical review platform requires the combination of features across existing software systems and

the implementation of new features that have not been included in any of the described systems.

CHAPTER 3

Classical ethical review and the Institutional Review Board process

3.1 History of ethical review

Even if researchers apply the strictest ethical standards to their work, it is still possible for their work to do harm, for example through instrumentalization in an unethical way by a third party or other inadvertent eventualities. It is often difficult to consider all possible ethical consequences even if researchers have the best intentions.[1]
Ethical review can assist researchers in conducting ethical research. The ethical evaluation of research projects is a rather new practice that emerged from incidents of unethical research which were largely due to malicious intentions. To better understand why IRBs today operate the way they do, this chapter first discusses how IRBs and the ethical review process in the U.S. developed. The second part then briefly describes the common ethical review process, called the "classical" ethical review process in this work. As will become clear, today's ethical review processes are closely connected to their historical development.

In the history of research with human subjects, incidents of unethical research including abuse and harm of research participants have highlighted the need for ethical oversight and review of research. Particularly horrific atrocities such as the Nazi crimes had been "disclosed in the Nuremberg military tribunal" [2, p. 1331] and led to the development and introduction of ethical standards in the field of medicine. In these standards it was stated, for example, that human subjects must consent voluntarily to research experiments [2, p. 1331]. Despite these terrible experiences other examples of unethical research studies and programs continued to occur even after the introduction of the Nuremberg Code [94, p. 311]. As a reaction to ethical violations in research, such as the Studies at the Fernald School where radioactive materials had been used in children in institutional care [94], [95] or the Tuskegee Syphilis Study conducted in rural Alabama between 1932 and 1972, in which individuals with Syphilis were denied treatment for observational purposes [96, p. 317] [66, p. 3], in 1974 the National Research Act was enacted, creating the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research and establishing the Institutional Review Board (IRB) system [94, p. 311f.]. The commission was tasked,

[1] For an example of such unintended ethical consequences see [93].

among other things, with the identification of basic ethical principles that should function as the basis for "biomedical and behavioral research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles." [3].

These principles and guidelines were published in 1979 in the so-called Belmont Report, on which IRBs are grounded to this day [3][97].[2]

The Belmont report identifies the following three general principles (with the remark that further principles may also be important) which should serve as an "analytical framework" [3] besides existing principles "generally accepted in our cultural tradition" [3] to aid researchers, human subjects, and reviewers, but also the general public, to better comprehend what ethical issues might arise in human subjects research [3]:

- Respect of persons; which includes both the autonomous treatment of individuals and the protection of vulnerable individuals
- Beneficence; which describes the obligation to do no harm and to maximize potential benefits while minimizing potential harms
- Justice; which refers to the duty to fairly distribute the benefits and burdens of research [3][94, p. 312]

From these ethical guidelines, the Belmont Report defines requirements that have to be considered in the conduct of research. The requirements are informed consent, the assessment of risks and benefits of research as well as the selection of human subjects [3].

While the Belmont Report remains the most important point of reference for IRBs today, it is also criticized that the regulations have not been updated since their introduction resulting in IRB "not align[ing] well with emerging practices and technologies" [9, p. 165]. The problems with current IRB processes will be discussed in chapter 4. In addition to and building on the Belmont report, the U.S. Department of Health & Human Services (HHS) together with the Food and Drug Administration reworked their existing human subjects regulations in 1981 [97]). In 1991 the Federal Policy for the Protection of Human Subjects, also known as the "Common Rule", was published "and codified in separate regulations by 15 Federal departments and agencies [...]. For all participating departments and agencies the Common Rule outlines the basic provisions for IRBs, informed consent, and Assurances of Compliance. Human subject research conducted or supported by each federal department/agency is governed by the regulations of that department/agency." [97] The "Common Rule" was revised and amended twice in 2018 [100].

IRBs inititally concentrated on biomedical research, but over time, their scope was expanded until nearly all research that involved human subjects was included by 1981 [101], [102, p. 314]. For example, IRBs were also created for social science research and research in the liberal arts [103]. Today, IRB approval is in general required in the U.S. by universities, funding institutions or scientific journals when a research project includes human subjects. In the U.S., the "Office of Human Research Protections, within the U.S. Department of Health and Human Services, is responsible for the registration of IRBs and their oversight" [103]. The aim of IRBs with regards to ethics can be summarized in protecting

[2] Besides the Belmont Report, other more specialized guidelines have also been published. One of the most important ones being the Declaration of Helsinki developed in 1964 by World Medical Association (WMA). The Declaration of Helsinki is a "statement of ethical principles for medical research involving human subjects, including research on identifiable human material and data" [98] and is directed mainly towards physicians. Later, the Good Clinical Practice (GCP) was introduced building on the Declaration of Helsinki to replace it [99].

subjects and enabling research beneficial to society [104, p. 2]. To conclude the history of ethical review in the U.S. and before moving on to the current ethical review process, it should also be noted that growing awareness of the diversity of sociocultural backgrounds including different moral but also religious values increasingly shows the complexity of human involvement in research but also stresses the importance of ethical review [105, p. 2].

3.2 The ethical review process in the U.S.

Current (or "classical") ethical review by IRBs follows a specific process which is outlined in this section.[3] Today, before a research study can take place, it has to be approved by an IRB [40, p. 5]. In 2012, about 4,000 IRBs existed just in the U.S. [40, p. 5]. Before describing the review process itself, three aspects of the organization of IRBs should be explained to better understand how they operate.

First, IRBs have to be independent from the investigator conducting the to-be-reviewed research as well as from the rewards of research [94, p. 313]. "The administration of an institution (e.g., president of a university or director of a hospital) must allow the IRB to function independently, without undue influence related to funding pressures or other administration priorities." [103] However, as [106, p. 19] argues, the understanding of independence varies, and it is not always clear how independent IRBs actually are. Second, and related to the first aspect, a distinction can be made between *local* and *central* IRBs. Whereas local IRBs are located at the academic institution where the research is conducted, such as universities, medical schools, hospitals and research institutes [104, p. 1], central IRBs are typically for-profit institutions that are paid by the researchers themselves [94, p. 312]. Both IRB forms come with their own challenges regarding independence (see for example [94], [106]). There also exist non-profit IRBs which try to overcome potential conflicts, but these are less common than central and local IRBs [106, p. 19]. Third, and finally, IRBs must be composed of members from different backgrounds: They have to include at least one member "from the scientific community" [103] as well as one member external to the scientific community. The latter's purpose is to "advocate for the nonscientific issues relevant to ethical conduct of research, such as legal issues and standards for professional conduct." [103] Moreover, IRBs must include an external member who commonly represents the community and "who may also serve as nonscientific member" [103]. For institutions that receive federal funding, IRBs have to consist of a minimum of five members and it is also possible for IRBs to invite consultants, who have an expertise in the research area of the proposal for additional information [40, p. 10], [103].

Before the beginning of the actual ethical review process, it has to be determined whether research involving human subjects is exempt[4] or must be reviewed by an IRB. This decision is commonly made by the IRB office or the Human Research Protection Program (HRPP) [108]. The Office for Human Research Protections (OHRP) and different academic institutions provide decision charts to determine whether a study is non-exempt human subject research [109], [110].

If research has been determined to be non-exempt human subject research, the researcher typically submits their proposal to the responsible IRB. In some cases, administrators assist

[3] For a detailed and thick description and analysis of the ethical review process see [40].

[4] This work focuses on non-exempt human subjects research. Therefore, details on the regulations and processes for exempt research are left out. It should be noted, though, that the Common Rule also describes a so-called Limited IRB review process for exempt research that collects personally identifiable and sensitive information (see [107].

the researchers with the requirements for application and take over a preliminary review of the documents for completeness before assigning the proposal to a reviewer [12, p. 4]. The proposal is then reviewed by the responsible IRB and either approved, rejected, or returned with change instructions [104, p. 1]. The review is usually discussed in a convened meeting of IRB members, which is called a "full board review", if more than half of all voting members of the IRB attend and of which at least one is a "nonscientist" [108]. The IRB's decision can be based either on a consensus approach or a majority vote. According to [103] "[m]ost IRBs use a consensus approach (i.e., votes must be unanimous) to reach a decision, although some IRBs allow a majority vote. When a majority vote is used, the community member typically still has substantial power because most IRBs will not override the perspective of the community member" [103].

Besides "full board review", the Common Rule defines another review process, which is "expedited review" [111]. Here, the review is conducted by one or several IRB members rather than by the full board. This makes the process more flexible and oftentimes faster. However, not all research can be reviewed in this procedure, it must meet certain requirements. In general, research may only include minimal risk and may not be invasive [112, p. 356]. They must also be covered by at least one of the categories specified in the HHS regulations for the protection of human subjects in research at 45CFR 46.110(b)(1)[112, p. 356] [113]. Moreover, while the reviewer is allowed to approve the research or to request changes to it, they are not allowed to reject it. If they consider the research to be ineligible for approval under the review criteria, the proposal is transferred to full board review [108]. The outlined IRB process must take place before participant or human subject recruitment ([103]. To be approved, the study protocol of the proposal must align with the principles described in the Belmont Report, as well as the Common Rule and institutional policy. In general, the principal investigator is responsible for the execution of the research [12, p. 3].

The IRB process does not end with the approval of research proposals. On the contrary, the IRB is responsible for overseeing the research which includes periodic "reviews to monitor the research progress and address any ethical issues" [103] and the possibility for research participants to contact the IRB if they have questions or concerns. The time intervals should correspond to the determined risk of the research but must be at least annual [96, p. 317]. If necessary, additional and surprise reviews are also possible [103].

Although IRB processes in the U.S. may differ slightly depending on the institutions, the described processes are well established and broadly applied. Yet, if we take a closer look, several problems of this IRB process become clear, ranging from regulations that do not match with new technologies, as mentioned above, to inefficiencies and questions of representation and independence. Moreover, there also exist problems that are not part of the IRB process itself, but that specifically emerge when we turn to areas of research beyond the medical field such as AI or to new scientific approaches to knowledge production such as CS. These problems of the current ethical review process in the U.S. will be subject of the following chapter.

CHAPTER 4

The need for a new ethical review process

4.1 Problems and challenges of the classical Institutional Review Board process

Due to the nature of its subject, ethics, IRB review is not an easy endeavor and includes challenges that cannot be resolved by simply changing the process. Whitney [104], for example, describes the inherent difficulty to weigh the benefits, risks, and expected knowledge of research based on the different principles of the Belmont Report which can push reviewers in different directions [104, p. 5]. These challenges will remain part of ethical review. However, the classical IRB process and its framing as described in chapter 3 have also been criticised by scholars and practitioners, who have highlighted several reasons that could be addressed to make ethical review more efficient, suitable, and equal. Without claiming completeness, 6 of the more significant problems that can be identified in today's IRB process that recur in the literature will be discussed in the following, together with issues raised in interviews with two IRB experts (for more information about the interviews, see chapter 6).[1] For further research discussing the current IRB system and its challenges see for example [115], [116]. This work does not deny the complexity of IRB processes and ethical review and does not disregard the efforts that have been put into existing processes. Instead, it aims at improving and facilitating ethical review by introducing a new adaptive and collaborative ethical review platform.

First of all, scholars observe how some of today's IRBs have lost their focus on ethics (see for example [94] and [117]). [94] states a mission drift of IRBs "that draws members away from the duty to ensure the fundamental protection of human subjects" [94, p. 315]. The author attributes this to two developments: the imprecise of definition of the terms "risk" and "research" which introduces vagueness into the review process (and its outcome) and overwhelming for IRBs and increasing concentration on documentation and the process itself which results in less time for "thoughtful review of important protocols" [94, p. 315]. This development could furthermore be linked to the growing number of research to be reviewed by IRBs over the years. To give a better idea of this growth, [118] conducted a comparison of new applications to be reviewed by federally assured IRBs in 1995 and 2005. While 105,000 new applications were to be reviewed in 1995, by 2005 this number increased

[1] Other issues that have been raised are for example the questioning of the actual independence of IRBs in [114].

to 181,669 new applications [118, p. 8]. Following, among others, [119], the authors argue that "[h]eavy workload may decrease review quality, increase research costs and negatively influence the willingness of faculty to participate on IRBs" [118, p. 10]. IRB expert 2, who was interviewed as part of this work, furthermore mentioned missing resources as cause:

> "I think this idea that IRBs are overworked, is true. I think, for in the United States, one of the reasons that's true is because the non-institutional, the independent IRBs are investor owned as we talked about, so they're all about cutting costs and constraining resources for economic reasons. And institutions IRB, people would rather spend the money on the research rather than the review, and so they don't get the resources they need. [...] I think that that's a bad perspective and it should change. They should have the resources they need to do it right." (IRB expert 2, June 7, 2022: 25)

Another reason for this perceived mission drift can be attributed to the increasing professionalization of university IRB staff since the 1980s which led to more focus on research governance and risk management [117, p. 3]. In a similar vein, Gunsalus et al. [120] have described the result of growing obligations of IRBs to review studies that usually include smaller risks for human subjects than physical or psychological examinations, including, for example, simple interviews or journalism, and the disproportionate paperwork as "simultaneous overregulation and underprotection" [120]. [2]
These concerns outlined in the literature were also raised and aptly summarized by IRB expert 1 in the interview conducted in the scope of validating the suggested CAER platform:

> "I think it's been reduced to a sidebar and what I mean by that is that I don't think that IRBs in the United States have, that's not their primary focus. Their primary focus is risk reduction, you know harm reduction. The regulations, I think, are seen as not something that's flexible, but something that has to be abided by very strictly. And again, that's part of that risk reduction, because they're afraid of their funding being lost and they're afraid of the FDA not approving something. And so I think ethics is really not something that anybody gives a lot of time to and I don't think it really comes into the discussion much at IRB meetings anymore. [...] I think, luckily, [the regulations are]] based on ethical principles, but [...] I don't think ethics is actively, an active concern." (IRB expert 1, June 1, 2022: 11-13)

IRB expert 1 further explained this change with the historical developments of the pharmaceutical industry in the US and growing research within the drug industry from the late 1970s:

> "And I think the drug industry, because it's largely not run by lawyers, but certainly, you know, the legal counsel in the pharmaceutical companies is very important. And I think when they started to do more research out in the communities and not just at academic institutions, you know, the the the focus was different. The focus became, you know, we're gonna make money on this drug. We have to do it quickly or we gonna lose money because patent laws and so everything just became this race to get IRB approval. And I don't think that even though we know that the ethics concerns are there, that I don't I think that the pharmaceutical company was able to just sort of, you know, Jam it down everybody's throats. And that was not the primary purpose or the primary concern.

[2] Similarly, other scholars have argued that today's IRBs have "lost the balance" regarding subject safety and public benefit and call for IRBs to find balance again and to "consider both the subject and society" [104, p. 2]. Whitney [104] also argues that today, IRBs sometimes focus more on protecting scientists from lawsuits than on protecting human subjects, and that this is due to the power attributed to IRBs [104, p. viii].

> It was about turnaround time of the IRB, you know, getting the approvals out. It was about how fast we'd bring this to market. So I think that ethical concerns became tied up in that that cycle of hush, rush, rush, rush get this done. And then, you know, the academic institutions that used to be concerned about ethical issues, they jumped on that that platform of, yeah, let's get this done, let's get some more money for our, our research and our funding. And so everybody just sort of jumped on what we call the bandwagon. And, you know, they just carried it out. And I think ethics was just left behind." (IRB expert 1, June 1, 2022: 17)

For these reasons, scholars and the interviewed IRB experts agree that IRB members should be able to focus on those protocols that are most important for ensuring the ethical conduct of research again and that the overhead of administration and documentation should also be decreased.

Second, another problem with classical ethical review is the composition of the review board itself, and the question of representation of the affected community. In most cases, IRB members are not sufficiently familiar with the local or affected communities and their values and norms [106, p. 18], and even though many academic IRBs include "community representatives", their involvement is oftentimes not meaningful as it is reduced to "simplify[ing] the language of consent forms" [94, p. 314] or because they feel subordinate to other IRB members [94, p. 314]. According to IRB expert 1, community representatives do not "really have much of a role at all." (IRB expert 1, June 1, 2022: 35)

Representation, and how it could be improved, is no easy question and there exists no general answer for it, since the answer not only strongly depends on the specific community, its structure, and values, but also on the particular research study, its scope, and how it affects the community and which community members in particular. I will return to the question of community representatives on IRB boards in the following section, since the role of community representatives only increases with the ethical review of new technologies and scientific approaches like HC and CS.

A third issue that has been raised in the literature is the imbalanced relationship between IRB members and the researchers whose work is to be reviewed. Originally, IRB communities had been imagined as "a panel of peers" [117, p. 9]. Oftentimes, however, the relationship between researchers and IRB members are far from collaborative. Thereby, according to [12], the discomfort exists on both sides: On the one hand, board members, who commonly volunteer their IRB activities, on occasion perceive that the researchers think of them as "obstacles to or adversaries in the research enterprise" [12, p. 5] whose effort and time are not valued. On the other hand, the researchers' perception of IRB members is that they are "unreasonable, obstructive" [12, p. 5] and only cause problems. At the same time, there is a clear power hierarchy between the researchers and IRB boards since no or only limited appeal possibilities exist for researchers [117, p. 9].

The interviewed IRB experts connected this gap to the problem of IRBs not being in service of the research participants but "more and more in service to sponsors and investigators and less and less in service to actual research participants. [...] And particularly when IRBs [...], the people, the groups that are supposed to be protecting research participants, are actually serving the researchers and the companies." (IRB expert 2, June 7, 2022: 13; see also ibid.: 5)

IRB expert 2 mentioned another reason for the gap between researchers and IRBs:

> "So the institutional IRBs here had developed a reputation of being very burdensome. They didn't meet frequently enough.[...] So the reviews were often sort of arbitrary, depending on what they remembered and who was in the room. They took a long time to approve protocols. They were going to review at once, ask questions. They reviewed the questions would be answered, they'd ask new questions that they should have asked the first time anyway. [...] So investigators really disliked them for good cause." (IRB expert 2, June 7, 2022: 13)

This observation leads to a forth point, which refers to the inefficient process of ethical review. In addition to repeated questioning and difficulties of reproducing what has been discussed in previous IRB meetings as described by IRB expert 2 above, completed reviews are usually stored away and cannot be reused. If new projects with the same ethical issues apply for IRB approval, it is not possible for reviewers to go back to a repository of previous research studies to read-up on similar research studies that have been reviewed in the past. The re-evaluation of every single issue is not only time-intensive but also makes it even more difficult to establish consistent evaluations in general and particularly for research fields that have only recently been added to those requiring IRB review [40], [15].

Following on from this fourth point, it is worth noting that the ethical standards of today's classical IRB processes, after having been tailored to specific fields, have not been updated since their first introduction and therefore may often "not align well with emerging practices and technologies" [9, p. 165]. IRB expert 2 described the emergence of new technologies that are not covered by current ethical review:

> "[T]here are a whole bunch of new technologies in citizen science with all of its dimensions. It's something that was never even envisaged when we wrote our laws. Never mind AI, machine learning, you know, genetic editing. All of that stuff is new. So there were all these new issues that it required sort of a new take on protections and potentially new regulation" (IRB expert 2, June 7, 2022: 13).

IRB currently apply "fixed principles from archetypal abuses" [104, p. viii] despite the changing nature of morality and the "more nuanced morality that leading scholars now propose" [104, p. viii]. Therefore, some scholars have requested a "reboot" of IRB [9]. Recent updates to the common rule have been made to attempt to better take into account emergent technologies and to "balance making the burden appropriate to the level of risk that people were put at while still effectively lessening unnecessary burden while still protecting people. I think in the end they heavily weighted towards lessening the burden." (IRB expert 2, June 7, 2022: 13) This led to "loopholes", like for example social media research which does not have to be reviewed because of the information on social media being considered public, which have not yet been fully addressed (IRB expert 2, June 7, 2022: 13).

Finally, the set of ethics applied in IRB are claimed to represent universal ethical guidelines. But as [121] shows with the example of the field of history, these ethics are in fact medical ethics that are imposed on non-medical research fields. Most of today's institutional or ethical review boards have been tailored to specific fields with a focus on biomedical and psychological research and experimentation [4, p. 115]. Considering the development of these boards, this, of course, comes as no surprise. However, this leads to problems in, for example, established sciences such as the humanities and social sciences as well as research fields in general where ethical review is now required but for which there exist no prevalent customized ethical review processes, such as CS, computing, and AI [66, p. 1].

For example, many scholars argue that IRBs oftentimes do not fit qualitative research due to a lack of knowledge about how qualitative research is conducted in general, its different epistemological understanding, or about the roles and relations of researchers and those researched which differ from biomedical experiments (see for example [122], [123], [124], and [117]. [117] explains: "Review boards, based on a medical model, can lack the philosophical understanding of qualitative research and therefore unnecessarily hinder research that lacks the clear boundaries between the so-called objective researcher and the subject" [117, p. 4]. With the advancement of machine learning and AI research over the last years, the problem of missing ethics becomes more and more pressing, as IRB expert 1 explained based on their experiences: "I've seen more and more studies come through on my desk for review about A.I. And there's no ethics in that. There's no concern about that at all in any of the applications that I see." (IRB expert 1, June 1, 2022: 19) This problem not only also applies to AI research and the field of HC-based CS, but is in fact especially pressing in these areas.

In the following, I will provide examples from this field to show where ethics in the field of HC-based CS diverge from ethics established in medical fields and inscribed in current IRB. Each scientific field has its own specificities, research approaches and practices which require thorough analysis and specific solutions. Whereas the dynamic ethical review system developed in this work aims at providing an adaptive framework that could be customized for various scientific fields, it will focus on the intersection of AI and CS to not remain in meta considerations, but to develop concrete ideas that could later be extended to other fields.

4.2 The example of human computation-based citizen science projects

HC-based CS projects are one example for new research fields and new technological applications which have only recently entered the view of IRB review in the U.S. and currently challenge some of the existing ethical principles of "classical" IRB review. They also raise questions about certain IRB procedures which are discussed in the following. In "the Atlas of AI" [4], Kate Crawford explains that for a long time "the fields of applied mathematics, statistics, and computer science had not historically been considered forms of research on human subjects" [4, p.115]. This position had been accepted by IRBs for many years, but with the trend of machine learning techniques and new focus on AI in recent years as well as the investigation of AI technologies from critical perspectives (e.g. social and cultural sciences, or science and technology studies), the sensibility for the impacts of these systems on individuals and society at large could no longer be dismissed. And with in some cases that are quite unlike those previously encountered, such as including humans in AI research by letting them collect training data for machine learning (ML) models, humans are in some sense central to the research without commonly being viewed as "subjects" in the traditional sense, thus raising questions about whether such cases are nevertheless to be understood as human subject research. One field in which HC applications are widely applied today is CS.

The field of CS itself (which——noted here for completeness——also goes beyond only HC-based projects), struggles with current IRB principles and processes. The focus in this work is primarily on CS projects that emerge out of the professional scientists' need for more computing power or human help for data analysis steps or data collection. But it should also be noted that CS projects can also emerge out of a community's interest and citizen scientists themselves. Here, project initiators sometimes face the problem that they cannot draw on IRBs of their institutions (as they may often not be directly affiliated with a

"traditional" institution like a major research university) and have to find commercial IRBs where the processes are not always transparent [125].

CS projects introduce new actors, relations and research practices into scientific projects that, to a large extent, have not been considered by institutional or ethical review boards before. At the same time, CS projects challenge some of the basic understandings of IRBs, such as the understanding that human subjects as participants in CS projects can sometimes be both researchers and subjects at the same time. Following [5], David Resnik states that "[t]he confluence of these two different roles in the same person poses challenges for investigators and oversight committees because legal rules and ethical guidelines focus on protecting the rights and welfare of human subjects and do not address issues that fall outside this domain" [6, p. 1]).

Another example of CS project values being inconsistent with the current IRB was described by Caren Cooper, Principal Investigator of the CS project *Crowd the tap1*, at the workshop "Toward reinventing IRB for Citizen Science" [126]: The idea of the CS project *Crowd the tap* was to protect the quality of drinking water. One of the outcomes of the project was that people would learn what risk they had of lead in water based on the types of pipes they had (Cooper at [126]). One of the "odd things in the consent form" [126] was that whereas the project understood this outcome as a benefit of the project, the IRB thought this would be a risk. This assessment might be related to the issue of "return of results", which has been strongly debated for many years (see for example [127]. However, it also shows how values that are common in IRBs focused on biomedical science do not apply to CS, where the results and the knowledge gained from the project can be one of the primary motivations of both participants and project designers. Besides these two examples, there exist a variety of other issues that arise when directly applying "classical" IRB values to CS projects, including, but not limited to issues around data openness vs. data protection [7], data ownership, and intellectual property [8].

The gap that exists in the ethical oversight in CS, according to [128], can be summarized as follows: "Citizen Science challenges virtually all of the conventional ethical gatekeepers: regulations, institutional policies, journal policies, funding streams, etc., and thus needs to approach ethics in its own terms" [128, p. 4]. Just as in the fields of computing and AI, there is a need for suitable ethical review in CS.

Taking the two fields of computing and CS and their ethical review problems as starting point, this thesis focuses on these fields' intersection in HC-based CS. Here, the ethical issues of both fields overlap in interesting ways that have not yet been addressed sufficiently in the discourse on IRB. This focus emerged from a collaboration I undertook together with the Human Computation Institute and our experiences with IRB applications for different projects. Specifically, for example, this work builds on our experiences with IRB processes for experimental studies in the HC-based CS projects *Stall Catchers* and *Dream Catchers*, showed that the IRB process could be more efficient both for the applicants and for the IRB experts. Under the current IRB process, applicants are required to answer all questions and their corresponding sub-questions, even if they altogether do not apply for the project under review. For example, it would require providing information on the research team's training in clinical procedures although the project was to take place not in any clinical context, but exclusively online. If only those questions that in fact apply to the to-be-reviewed project's context and field were posed to project representatives, the IRB process would be significantly less time-consuming and more efficient overall.

Most importantly, however, these experiences and the discussion of the question of

ethical review for online CS in the ECSA2020 workshop [126] made clear that the biggest problem for the ethical review of HC-based CS projects today is ultimately the lack of ethical guidelines for this field. As a very new field, ethical implications, norms and values are not yet established, which implies that no specific ethical guidelines or principles that can be applied in ethical review of HC-based CS projects actually currently exist.

IRB expert 2 similarly argued that, currently, such research is reviewed by reviewers without a particular expertise in these new emergent fields. For example for a biomedical HC-based CS project that includes AI, biomedical researchers and doctors might review the project who might

> "know a lot about the benefits [of machine learning and AI]. They know what they want to learn. But they don't know particularly about the harms, you know, the group harms, the biases, the. And so really, they shouldn't have authority. You know, you shouldn't have authority if you can't balance that. [...] I think these we have to be very careful about these these issues, about where your authority comes from and who gave you that authority are really important." (IRB expert 2, June 7, 2022: 13)

This is why in this new field, it is even more important to include the participants and affected communities not only into the ethical review of such projects, but also into the development of the actual ethical guidelines for this field. Given that today's IRB committees mostly include only one community representative, who incidentally also does not necessarily have a lot of decision power in the process as discussed in chapter 4, it becomes clear that HC-based CS poses a significant challenge to current IRB.

Due to the lack of understanding of moral values in the field of HC-based CS, and consequently the lack of ethical guidelines for this emerging field, the challenges to the IRB from HC-based CS mentioned here represent just a few examples of an area that still needs to be explored. To advance and facilitate this process and to solve some of the major problems the IRB process faces today in general and especially in the field of HC-based CS, this work suggests a new approach towards collaborative and adaptive ethical review of HC-based CS with a new system called CAER. This approach will be introduced in the next chapter.

CHAPTER 5

CAER - a collaborative and adaptive ethical review platform

This chapter introduces the new collaborative and adaptive ethical review system, called CAER. This work develops further the initial ideas for a new approach to ethical review described in [15]. It should be noted that this work does not aim to build or define ethical consensus among the involved stakeholders as such, but rather to better enable information processing and collaboration on a platform in order to improve the ethical review processes in general. Although the system is adaptable to all different kind of research fields, the proposed system especially focuses on the intersections of AI-research and CS with HC-based CS projects to provide a specific context which allows to test and evaluate the system.

The envisioned CAER system is a web-based platform to facilitate the ethical review process for trustees (IRB experts and community representatives) and project representatives (researchers with a project to be reviewed) and allows to adapt underlying ethical guidelines with the help of HC to make the review process more suitable for specific research fields.

In the following, first, user stories for the main two stakeholder groups and roles in the new review system are described. Building on these, I then outline the main functionalities of the envisioned system with workflows. Thereafter, the technical specification is described before finally visualizing the workflows in process diagrams and discussing further aspects of the design of the CAER system.

5.1 Design and Requirements

5.1.1 User roles and personas

There exist two main user roles in the new review system: The first are **project representatives** of a new HC-based online CS project who would like to let their project be reviewed before its launch. The second are so-called **trustees**. This is a heterogeneous group of individuals that includes, for example, IRB experts, researchers and community representatives, who are trusted members of the ethical review and scientific community or the group of people that are affected by new HC-based CS projects. For the sake of simplicity, and because they play the same user role in the review system, this heterogeneous group

of people will be referred to as "trustees" in general. The main activities of trustees, in the context of CAER, is to manage ethical guidelines and perform human review of applications where needed.

For the design and development of new software systems so-called *personas*, imagined representations of user roles have been proven to be very helpful [129, p. 58f.]. The following table describes three personas, each representing an important user role of CAER.

Name and Role	Detail	Goal
Kim, project representative	A researcher who has been working together with their team of four colleagues to build a new online CS game that builds on HC to solve their current data analysis problem. Kim has never build a HC-based CS game before.	Would like to undergo ethical review with their new project before launching it.
Irene, IRB expert (trustee)	An IRB reviewer with many years of experience for whom ensuring ethical and responsible research is a matter of the heart but who is sometimes overwhelmed with the large number of applications.	Would like to carry out the ethical review conscientiously and carefully.
Lars, community representative (trustee)	Father of two children who works in a medium-sized company in Cleveland and lives in a suburb. In his leisure time he enjoys contributing to online CS projects, especially those that focus on environmental issues in his surroundings.	Meaningfully contribute to research that could have a positive impact on the environment and his family's conditions of living; to contribute to keeping ethical guidelines up-to-date so that he and his fellow citizen scientists are protected and well represented.

These personas help make the user stories presented in the following more meaningful and build a basis for discussing the new CAER system in its different development stages [129, p. 58]. A full implementation of the CAER system is not developed in the scope of this work, nevertheless the personas and user stories may also facilitate future research and work on this new system.

5.1.2 User stories

This section describes the *epics* and *user stories* for the new CAER system. User stories are a helpful planning tool, since each user story refers to a specific functionality, meaning something a user would perform in a given situation [129, p. 36]. The user stories outlined in the following build on the above described personas and follow the properties defined by Bill Wake [130] and discussed by [129]. These properties can be summarized with the abbreviation INVEST: A user story should be **independent** (they should not be dependent on other user stories), **negotiable** (user stories are not requirement specifications but support discussion), **valuable** (for users and/or customer), **estimable** (developers should be able to estimate scope and effort needed to implement story), **small**, and **testable** [130]. Epics, by comparison, are somewhat more extensive and more general, and are therefore well-suited

as a high-level point of departure. Therefore the epics are described first, and subsequently specified into user stories, for both user roles.

These epics and user stories have been derived to varying degrees from user and expert interviews (see chapter 6), secondary research (see chapter 2) as well as my own first-hand experience with the ethical review process with IRBs in the US.

The first epic to be described refers to the user role of a project representative such as Kim:

> As project representative Kim would like to know if their new HC-based CS projects fulfills current ethical guidelines.

User Stories for project representatives:

- 1a) Kim wants to enter their project details so that it is simple and it does not cost them too much time.
- 1b) Kim wants to be able to follow the decision-making process so that they can improve their project to meet the requirements.
- 1c) Kim wants to be able to discuss open questions with the reviewers to make sure they understand the reviewers' reasoning.
- 1d) Kim wants to be able to appeal so that they can express their thoughts if they disagree with a decision.

The second set of epics refers to the user role of a trustee and can be divided into a individual epic each for Irene and Lars:

> As trustee Irene would like to carry out the ethical review conscientiously and carefully.

User Stories:

- 2a) Irene wants to have all relevant information about a project available so that they can carefully evaluate the project on the basis of current ethical principles.
- 2b) Irene wants to spend as little time as possible with the review of already known issues so that they can focus on new issues for which there exist no ethical answer.
- 2c) Irene wants to be able to adapt ethical issues or rules if they are outdated or do not fit current ethical understandings.

> As trustee Lars would like to contribute to the establishment and maintenance of ethical guidelines to ensure these are suitable for the field of research the CS project he participates in and that his participant voice is being heard.

User Stories:

- 2d) Lars (as well as Irene) wants to be able to add new ethical issues if ethical problems arise in projects that have not been considered in the ethical principles so far so that they can ensure that the ethical principles applied to the review of new projects meets are suitable.

- 2e) Lars wants to review new suggested ethical issues and rules before they are included in the ethical guidelines to make sure the community's values are being considered.

These user stories effectively correspond to the main functionalities the CAER system should have to meet the needs of the user roles. From these, it is now possible to derive the requirements and specifications for the new ethical review process with CAER which will be described in the following section.

5.1.3 Workflows for ethical review with CAER

At the core of CAER are the following four main workflows:

- Project submission and automated review
- Manual project review
- Managing ethical guidelines – change requests
- Managing ethical guidelines – change approval

Additionally, miscellaneous workflows such as user management and notification may also be needed for a "real-world" implementation of such a system, but for the sake of brevity these are not further described here. In the following, each workflow is described to provide a better understanding of the envisioned system and its main functionalities. On this basis, further technical details are subsequently be described in the functional specification and requirements in section 5.2. The described system aimed at is a "minimum viable" form of the envisioned collaboration platform that focuses on the main system functionalities for the review of HC-based CS projects. "Minimum viable" in this context simply means that the system (or rather, its description) should capture all core required workflows to make an evaluation and limited practical application possible, but that for large-scale use in the real world, additional features and specification may be required that is omitted in this work.

5.1.3.1 Workflow 1: Project submission and automated review

Kim, who wants to undergo ethical review before conducting a research project or study registers to (or, if they have used the platform before, logs in) to the platform as project representative. When starting a new review process, they are presented with a form of questionnaire in which are asked to first provide information on their project and then to respond to questions. One question is presented at a time, and the next question is subsequently presented depending on the specific response provided and the importance of specific ethical issues to the research domain and form of the project[1].
Once all relevant questions have been responded to by Kim, the outcome of the automated ethical review is presented. Kim's project is either accepted under the current ethical guidelines, or it is rejected if the project does not meet the requirements for ethical approval. Both results are preliminary, which means that projects will only be finally accepted or rejected after a second review by a human expert.
Project representatives have the possibility to request more information on the decision process. In this case, the history of their question-response flow is displayed, with additional information on each question-response pair. For example, for the question of "Does the project collect and store the date of birth of users?", an information box may explain that it is currently understood to be problematic to store the date of birth of users, since

[1]The underlying customized rule engine is described in subsection 5.2.4.

this forms sensitive personal information and that instead, it would be better to store the current age, which ensures the privacy of the users. With this information, project representatives gain a better understanding of ethical issues (educational value). If their project is rejected by the automated evaluation, project representatives like Kim can decide to revise their project based on the provided information and re-submit it at a later time, or, if they do not agree with the outcome and information provided on a certain question, or believe that certain questions do not apply to their project, or that some answer-possibilities are erroneous or incomplete, they can request the modification of existing questions or suggest new ethical issues to be considered in the future review process. This is known as an *appeal*, a functionality that is envisioned for CAER but not commonly found in existing review systems/processes. Such feedback is sent to the trustees, who then review it and, if they believe it to be valid, incorporate the feedback into the system (up to the point of in fact reworking existing ethical guidelines and/or the rules with which they are applied).
Once the automated evaluation is complete and the project representative has not decided to first revise the project and to re-submit it at a later stage, the project information together with the history of question-response pairs is forwarded to selected human reviewers.

5.1.3.2 Workflow 2: Manual project review

Trustees, such as Irene, who are registered to the platform receive notifications about the submission of a new research project after they have been automatically evaluated by the system. If she accepts to review a specific project, she is presented with the project information and the history of ethical questions-response pairs provided by the project representative via the CAER platform. She can now proceed with the review of the project. Since most known and relevant ethical issues have commonly already been addressed by the system, she can simply read through these and then focus on potential open questions or issues that have not been addressed by the system. Trustees provide written argumentation on the project and their decision to accept or reject the project via the platform. Once a final decision is made, project representatives are informed via the platform.

5.1.3.3 Workflow 3: Managing ethical guidelines – change requests

To keep the ethical guidelines underlying the automated evaluation up-to-date, trustees, such as Irene and Lars, can request updates to existing ethical questions or rules or can suggest new questions to be considered in the automated evaluation together with new rules. For this purpose, trustees first choose whether they want to either make a change to the ethical *questions*, i.e. suggest a new ethical question or update existing questions, or whether they want to make a change to the *rules* that specify how the responses to ethical questions are applied to projects.
If trustees want to update an existing ethical question, they can do so by choosing the question from the list of all questions (a search option facilitates finding the question). Updates can be made by editing the question. All changes receive the status "pending for review" in order to be tracked and presented to other trustees for review. By providing additional information on the changes made trustees can explain their motivation to update an existing question.
To suggest a new question, trustees can simply type the new question in a text field and provide additional notes in an information box. Once they've submitted a new question, the new question received the status "pending for review" and other trustees will be notified and requested to review the new suggestion (see workflow 4). Once they have suggested a new question, trustees will be asked if they would also like to add a new rule related to the new question. Note that in this initial design on CAER, like in classical ethical review,

only questions that accept yes, no or N/A as an answer can be used. However, this is not a fundamental property of CAER and could relatively easily be extended to allow arbitrary multiple-choice, possibly also numeric or other modalities of answers. In addition, project representatives have the opportunity to textually explain any potentially relevant context or nuance around their response that reviewers should consider.
Updating rules follows a similar process. To update existing rules, trustees select the desired rules from the list of existing rules (a search option facilitates finding the desired rule). They can then directly modify parts of the rule such as exchanging the response values to a certain question or exchanging a question if the rule consists of two different questions and corresponding response values. They can also update the operator between the individual rule elements.
To suggest a new rule, trustees have several possibilities to choose from:

- Trustees can choose the option to create a rule from an existing question and specific response values. In this case, they can choose from the list of existing questions and add one or more desired response values. Before submitting the new rule, they can select, using a checkbox, if the new rule leads to rejection of the to-be-reviewed project (this is needed so "combined" rules can be created without each "sub rule" by itself already leading to rejection).

- Trustees can choose the option to create a rule from two existing questions. This process is similar to the first case, but for two questions. Additionally, trustees have to choose a boolean operator ("And", "Or") to combine the responses to questions. Before submitting the new rule, they can again select whether the new rule leads to rejection of a research project.

- A new rule can also be created by combining an existing rule with another question-response pair. In this case, trustees first choose the existing rule from the list of all rules and then select a question and the desired response values. These elements are then combined with an operator before submitting the new rule. Again, trustees can select via a checkbox whether the new rule leads to rejection of the project.

- Finally, trustees can create new rules by combining two existing rules by selecting the desired rules from the list of existing rules and combining them with the "And" or "Or" operator. A single question or rule can also be negated using a "Not" operator.

All of the above described cases include a text box to provide additional information on the motivation and reason to suggest a new rule.
By providing different options to create a rule but at the same time restricting the process to one or two elements to combine, this workflow remains easy to understand and ensures usability. Due to the clear action options this also reduces sources of error. If necessary, more complex rules can be created by repeating the workflow. The fact that rules can be composed, i.e. rules can reference other rules, makes it possible that even more complex requirements can be expressed without making any individual rule overly complex to author or interpret.
All changes to the rule collection are suggestions, not direct edits. All changes are tracked and set to "pending for review". After submission, notifications are sent to other trustees for discussion and review of the changes. If the changes are accepted, they are subsequently used in any further automated review of submissions (but not retroactively applied to previously submitted projects).

5.1.3.4 Workflow 4: Managing ethical guidelines – change approval

If new change requests to existing ethical questions or rules are submitted, or suggestions for new questions or rules are made, three trustees[2] will automatically be notified and asked to review the suggestions. Trustees can then review the changes and suggestions on the interface provided for this purpose by choosing one of the options "accept", "reject", "accept under following conditions" and provide explanation as well as, if applicable, change suggestions in a text box. Trustees can then discuss the changes and come to an agreement. If all trustees accept a change or suggestion, it is included in the current set of ethical issues, meaning the questions and rules used for evaluation of new submissions.

5.1.3.5 Miscellaneous workflows

The focus of this work lies in developing a new HC-based ethical review process that improves current IRB processes and allows adaptive and more suitable evaluation criteria (ethical guidelines) for new emergent research fields such as HC-based CS. However, additionally to the above described core workflows, miscellaneous workflows such as account creation, user management, communication, platform administration, logging, and notification are also necessary. These workflows do not have requirements that are specific to the envisioned CAER platform but are similar to other web applications. Although these functionalities are very valuable and important for such a collaboration platform, there already exist ready-made solutions that can be integrated into the system at a later stage (see chapter 2).

5.1.4 Miscellaneous requirements

Before turning to the technical specification three additional requirements should be noted. First, since different users should be able to access the platform from different places and at any time, the platform will be web-based.
Second, to make the platform user-friendly and accessible for different user groups with potentially no technical background, it is important to ensure a purely text-based implementation of the system.
Finally, platform development should follow an open-source approach. Not only has the review of existing IRB software made clear how difficult it is to understand the review processes and flows from the software's description if the software is proprietary, but the importance of transparency has been stressed by the interviewed experts. To build trust in the review processes and decisions, transparency has been described to be essential by the interviewed IRB expert 2 (IRB expert 2: 13–15). Open-sourcing the software is one important step into this direction, which can additionally increase resilience of the platform.

5.2 Technical specification

Having described the design and requirements of CAER, in this section the platform is further specified on the technical level. First, the system architecture, which defines interaction between application layer, data model, and business logic, is discussed. Then, the data model is described and specified with figure 5.2, followed by the specification of the

[2]For the minimal viable version of the platform, the number of trustees that have to agree to ethical guideline updates is set to three. This means that in total, four trustees have to agree that a certain change is desirable – one trustee requesting the change and three trustees reviewing it. In future, to make the system as efficient and effective as possible, the required number of trustees should likely be derived through experimental testing (see section on future work 7.4).

rule engine. Last but not least, to bring together the design, workflows and technical specification, process diagrams describes the logical flow of the core processes followed within the application.

5.2.1 System architecture

The system architecture of the CAER platform as displayed in figure 5.1 consists of three major parts. These are the backend with its rule engine and data layer, the application layer, which describes the frontend, and data storage.

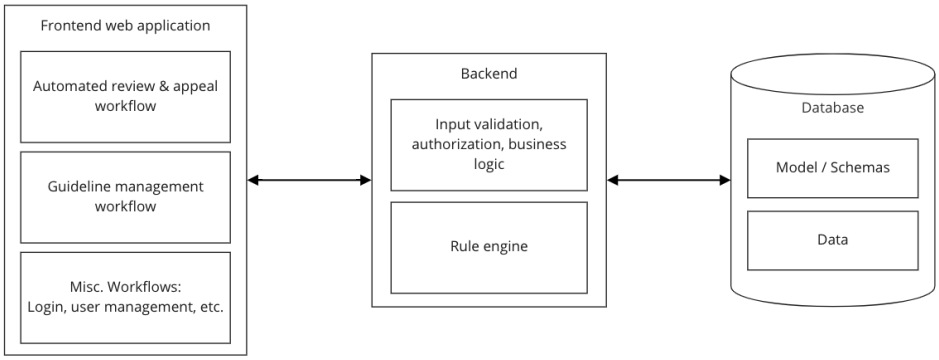


Figure 5.1: CAER system architecture

The frontend determines how users interact with the system and what they can see. It communicates with the backend via a JSON-based Application Programming Interface (API). It consists of 6 main user interfaces:

- A start page with information on the CAER platform and features such as registration and login options,
- a project representative managing view where researchers can submit their research projects for ethical review, keep track of submitted projects, and communicate with trustees,
- a automated evaluation form view where the initial evaluation of research projects takes place,
- a trustee evaluation view, in which trustees such as IRB experts review and discuss projects and communicate with project representatives,
- a trustee ethical guideline view for managing the underlying ethical guideline by updating existing and/or suggesting new ethical issues and rules, and last but not least
- a repository view which includes a history of all reviewed projects and and reviews that can be searched.

The backend is in control of all interactions with the data and enforces that "business logic" is correctly applied, i.e. that no illegal or nonsensical actions. This also means it ensures that only valid interactions and data are written to the database, since the frontend could be manipulated or bypassed by malicious users/actors in a web-based application. Since it describes one of the core aspects of the envisioned CAER platform, the rule engine

included in the backend is explained in detail in section 5.2.4.

As has been established as good practice, the user interface is decoupled from the backend to facilitate collaborative software development and maintenance. Communication between frontend and backend is therefore managed through stateless APIs. The business logic forms a mediator between the APIs and the data model. Data, whose properties and relations are defined by the data model, are stored in a relational database. This implementation pattern is ubiquitous in web applications today and can be seen as an implementation of the Model-View-Controller (MVC) pattern, first described at least as early as 1988 [131], but, as a term, commonly applied to this style of web application today [132]. Roughly speaking, in this framing, the data layer and database corresponds to the Model, the backend represents the Controller, while the frontend can be seen as the View. More correctly perhaps, the backend and frontend could (or should) be seen to some extent as a joint implementation of View and Controller (and internally, the frontend also manages model-like representations), but the emphasis here is on the separation of "business logic, presentation logic, and request processing" [132] in much the same way as [132] apply the term.
In the following section, the data model is described in more detail.

5.2.2 Data model

Figure 5.2 summarizes the data model (or schema) in a class diagram with data types and relations between the different classes. To keep the focus solely on the data model, potential methods of a final implementation have been omitted from this class diagram. Similarly, models needed for change management and approval processes have been omitted to focus on the core information architecture. The design of this data model has been inferred by careful analysis the user stories and workflows previously described.

The data model consists of the following classes:
Project, which is used to describe research projects or studies that have either been submitted to the platform (status "under human review"), are currently in submission, which means that the automated evaluation has not been completed yet ("pending"), or have undergone automated and human review and have either been approved or rejected.
The class **User** describes a user of the platform who can have the role of a trustee or a project representative. Each projects refers to the project representative(s) that has/have submitted the project and trustee(s) who review(s) it.
The class **Question** includes the individual questions generated from the ethical guidelines and is connected to **Response**.
Response describes the responses to specific questions provided by the project representative for a specific project. In accordance with current IRB forms, response options (ResponseValues) are "Yes", "No", or "N/A" (Not Applicable).
Finally, a **Rule** is a connection between either a) a question with a specific response values, b) two questions with specific response values each, c) an existing rule and a question with a specific response values, or d) two different existing rules. The operators between the elements of b) to d) can be the logical "And" or "Or".
Users are only allowed to either

- fill a question-id and response values
- or rule-ids and operators.

The operation type "Not" may only be used for a single rule, i.e. it is a unary operator.

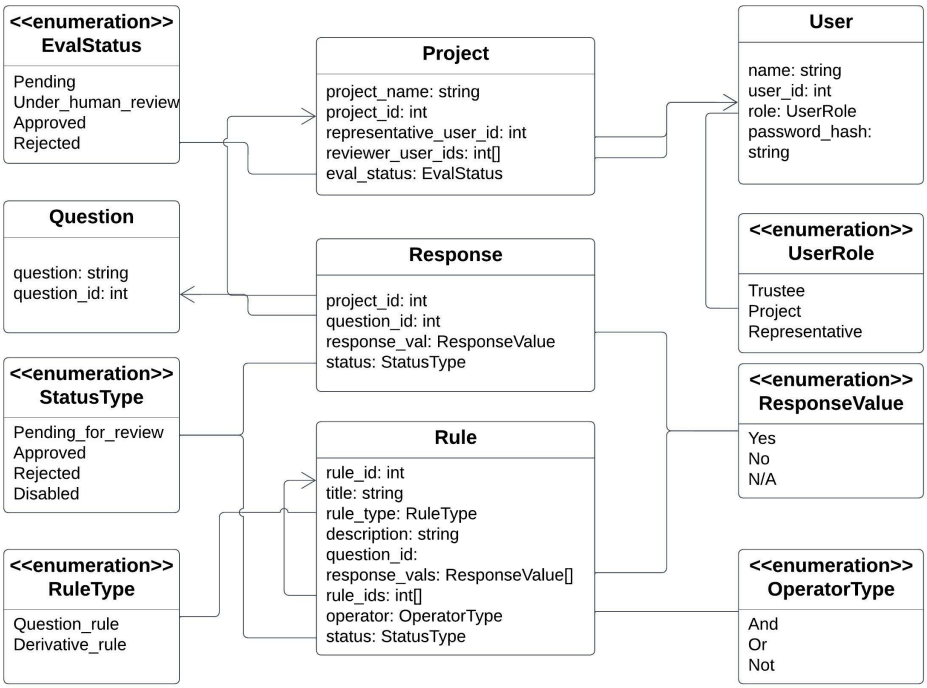


Figure 5.2: Data Model UML class diagram

5.2.3 API Specification

The following section provides possible API specifications using JSONSchema[3] format for input and output data models. In the following the APIs for project management, question management, and rule management are described in individual subsections. Input and output schemas are listed for each route where applicable (some routes may return only integers or require no inputs). In a final implementation, all APIs should be authenticated and access controlled for the current user, but details of such an authentication/authorization scheme are not reflected in this specification. The miscellaneous workflows may also require further specific APIs which are omitted from the following.
APIs are required for project management, question management and rule management. All endpoints for the individual APIs are defined in the following subsections.

5.2.3.1 Project management API

The project management API is an HTTP-based JSON service that is used to create, read, update and delete projects throughout their lifecycle. It creates a new project, stores them to the database, and returns the created project data in JSON format.

Endpoints for the project management API are described in the following tables[4].

[3] `https://json-schema.org/` [Accessed: Sept. 11, 2022].
[4] In order for the specifications to fit on the pages, references to named schemas via $*ref* were not always resolved. A full list of all schemas can be found in appendix A.1

Path	/projects/
Description	Fetches a list of all projects this user is allowed to see.
Method	get

Output schema	{
	"type": "array",
	"items": {
	"$ref": "#/components/schemas/Project"
	}
	}

Path	/projects/
Description	Create Project
Method	post

Input schema	{
	"required": [
	"name",
	"description",
	"user_id"
	],
	"type": "object",
	"properties": {
	"name": {
	"title": "Name",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"user_id": {
	"title": "User Id",
	"type": "integer"
	}
	}
	}

Output schema	{
	"$ref": "#/components/schemas/Project"
	}

Path	/projects/{project_id}/response
Description	Submit Response
Method	post

Input schema	{
	"required": [
	"project_id",
	"question_id",
	"response_value"
	],
	"type": "object",
	"properties": {
	"project_id": {
	"title": "Project Id",
	"type": "integer"
	},
	"question_id": {
	"title": "Question Id",
	"type": "integer"
	},
	"response_value": {
	"$ref": "#/components/schemas/ResponseValue"
	}
	}
	}

Output schema	{
	"$ref": "#/components/schemas/Response"
	}

Path	/projects/{project_id}
Description	Update Project
Method	put

Input schema	<pre>{
	  "type": "object",
	  "properties": {
	  "name": {
	  "title": "Name",
	  "type": "string"
	  },
	  "description": {
	  "title": "Description",
	  "type": "string"
	  }
	  }
	}</pre>

Output schema	<pre>{
	  "$ref": "#/components/schemas/Project"
	}</pre>

Path	/projects/{project_id}
Description	Delete Project
Method	delete

Output schema	<pre>{
	  "type": "integer"
	}</pre>

5.2.3.2 Question management API

The question management API is an HTTP-based JSON service that is used to create, read, update and delete questions throughout their lifecycle. It creates a new question and returns the created question as JSON.

Endpoints for the question management API are described in the following tables.

Path	/questions/
Description	Fetches a list of all questions.
Method	get

Output schema	<pre>{
	  "type": "array",
	  "items": {
	  "$ref": "#/components/schemas/Question"
	  }
	}</pre>

Path	/questions/
Description	Create Question
Method	post

Input schema	{
	"required": [
	"question"
	],
	"type": "object",
	"properties": {
	"question": {
	"title": "Question",
	"type": "string"
	}
	}
	}

Output schema	{
	"$ref": "#/components/schemas/Question"
	}

Path	/questions/{question_id}
Description	Update Question
Method	put

Input schema	{
	"type": "object",
	"properties": {
	"question": {
	"title": "Question",
	"type": "string"
	}
	}
	}

Output schema	{
	"$ref": "#/components/schemas/Question"
	}

Path	/questions/{question_id}
Description	Delete Question
Method	delete

Output schema	{
	"type": "integer"
	}

5.2.3.3 Rule management API

The rule management API is an HTTP-based JSON service that is used to create, read, update and delete rules throughout their lifecycle. It creates a new rule and returns the created rule as JSON.

Endpoints for the rule management API are described in the following tables. As can be seen from these tables, the rule management API also constraints the possibilities for users to create rules as has been defined in the data model to make sure that only valid rules can be created.

Path	/rules/
Description	Fetches a list of all rules.
Method	get

Output schema	<pre>{
	  "type": "array",
	  "items": {
	  "$ref": "#/components/schemas/Rule"
	  }
	}</pre>

Path	/rules/
Description	Create Rule
Method	post

Input schema	<pre>{
	  "required": [
	  "rule"
	  ],
	  "type": "object",
	  "properties": {
	  "rule": {
	  "title": "Rule",
	  "anyOf": [
	  {
	  "$ref": "#/components/schemas/QuestionRuleCreate"
	  },
	  {
	  "$ref": "#/components/schemas/DerivativeRuleCreate
	  "
	  }
	  ]
	  }
	  }
	}</pre>

Output schema	<pre>{
	  "$ref": "#/components/schemas/Rule"
	}</pre>

Path	/rules/{rule_id}
Description	Update Rule
Method	post

Input schema	{
	"required": [
	"rule"
	],
	"type": "object",
	"properties": {
	"rule": {
	"title": "Rule",
	"anyOf": [
	{
	"$ref": "#/components/schemas/QuestionRuleUpdate"
	},
	{
	"$ref": "#/components/schemas/DerivativeRuleUpdate
	"
	}
	]
	}
	}
	}

Output schema	{
	"$ref": "#/components/schemas/Rule"
	}

Path	/rules/{rule_id}
Description	Delete Rule
Method	delete

Output schema	{
	"type": "integer"
	}

5.2.4 Customized rule engine

Since the rule engine forms one of the core aspects of the CAER platform, a minimum version of the required rule engine that fulfills all necessary but only those necessary requirements for the review system is described in this section.

The rule engine is key for ensuring effective and efficient project review. It evaluates which questions should be presented to the project representative and in what order. As became clear in chapter 4 a current problem of ethical review of HC-based CS projects is that they are evaluated on the basis of a catalogue of question which had been developed for different fields and therefore may include questions that not not apply to HC-based CS and at the same time miss others that could be relevant. To make the ethical review process more efficient, only those questions should be asked that are relevant to a certain

project. Thus, the rule engine should evaluate after each question which question to pose next based on the last answer provided by the project representative. Predicting or inferring the next question is therefore one of the core functionalities of the suggested adaptive system. An early consideration in the design of the rule engine component was to rely on existing rule engine solutions or even logic-based programming languages such as *Prolog*, as the design of these systems usually assumes the presence of all responses before the evaluation of a given rule, rather than being designed towards interactive selection of the "most efficient" question to pose next, this approach did not turn out to offer tangible benefits over one where the rule engine of CAER is custom-built according to the criteria outlined here. It is also the case that, for prototyping such a rule engine, a simple smart form design is sufficient (even though this requires more manual work in setting up the rule conditions than a full implementation might).

The purpose of the rule engine is to evaluate a project according to the underlying ethical issues (questions & answers) and rules. This is achieved via the following basic process and conditions:

- Each question and response pair is individually posed to the user and immediately evaluated, so that according to the last question and response pair, the ideal next question will be posed (and has to be answered by the project representative). "Ideal" here can refer to different objectives, but for simplicity it can be assumed that the ideal next question is the question with highest expectation that the project representative's response may lead to the project being rejected. Similarly, an "ideal" question is always a question that actually needs to be answered at all - if the question/response pair in question is only used in a rule (or multiple rules) that logically cannot be fulfilled anymore based on previous responses, that question should not be posed to users at all.
- Implied in this is the fact that all submitted rules in the system correspond to a condition under which the outcome "further review necessary/ denied" would be assigned in an evaluation. This is to make the goal of the system well-defined ("pose those questions first that are most likely to lead to rejection"), as well as to prevent contradictions to be introduced into rule system by its users, as a mixture of positive and negative rules is unnecessarily confusing and may easily lead to incorrect states.
- Trustees therefore only add rules describing problematic submissions, meaning combinations of issues that would resolve in "further review necessary/denied" or "connection" rules which are meant to form parts of more complex rules that lead to rejection of a project.
- Correspondingly, if all rules are unfulfilled (which generally does not require users to answer all possible questions, only some set of questions that can be inferred to make all rules impossible to fulfill), the project is evaluated as acceptable (pending a final human confirmation).

The following example further illustrates how rules can be added by trustees to ensure the rule engine only has to follow the solution "further review necessary/ denied":

Given the following ethical issues:
I: Issue "collect personal information" with answer "yes"
II: Issue "personal information will be anonymized" with answer "no"

A following new rule could now be introduced:

I AND II → "further review required/project rejected"

The following forms a counterexample for a new rule that would not be accepted by the system:

I: Issue "collect personal information" with answer "yes" II: Issue "personal information will be anonymized" with answer **"yes"**

The following rule would not be directly expressible in the system:

I AND II → "project approved"

Instead, the following rule could be written:

(NOT I) OR (NOT II) → "project rejected"

In classical ethical review, questions can implicitly have subquestions, but even if the "parent" question implies a certain response to the subquestion, both responses are usually required. In CAER, subquestions are not explicitly formulated as such but rather result from a rule combining the "parent" and "child" question. To ensure that such sub-question are not asked before their corresponding parent question, trustees may need to specify the order of such pairs of questions, as that order cannot necessarily always be inferred through the conditional structure itself. For example, if question B is a subquestion of question A, and either response to question A is acceptable as long as the response to B is negative (but negative response to A implies that the response to B must be negative anyway), in CAER there would be a rule *(NOT A) AND (NOT B) → "project rejected"* which sufficiently captures the negative outcome. For simplicity and assuming no other factors influencing this scenario, it can be assumed that CAER will ask A first simply because it occurs before B in the above rule (but if a positive response to A is given, the entire rule can be assumed unfulfillable and question B is not asked). Future work could include investigation of more advanced approaches, such as statistical analysis that helps determine which question between A and B is expected to lead to a "quicker" evaluation based on past responses.

As has been described in chapter 2, there exist so-called "smart form builders" that allow to create basic versions of rule engines with conditional logic. For better understanding of the rule engine and for evaluating its functionality, a proof-of-concept prototype has been developed with *Airtable* and will be described in chapter 6. As will become clear, existing platforms have limitations that do not allow to develop the CAER platform with all its functionalities. However, they are still a good starting point for validating the idea.

5.2.5 Workflow process diagrams

To bring together the design of the platform, its workflows, and the technical specification, the following workflow process diagrams (5.3, 5.4, 5.5) describe the main information and process flows of CAER. In these diagrams, swimlanes represent users (via their roles) using the system as well as the CAER system itself.

5.3 Discussion of CAER design

In this chapter, the CAER platform has been introduced with regards to its design, the main technical requirements, and its technical specification. Compared to existing ethical review

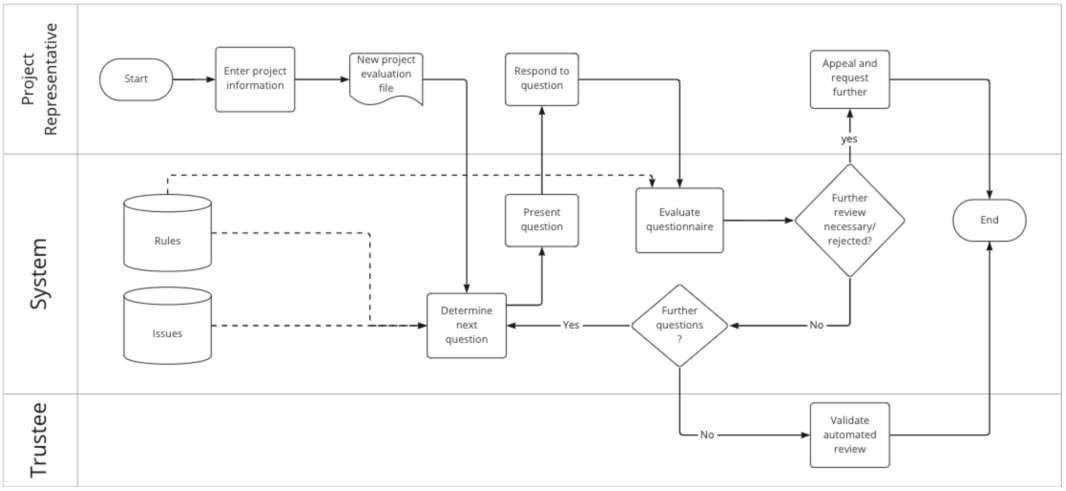


Figure 5.3: Process diagram for project evaluation

system platforms, CAER allows for dynamic adaptation to the specific requirements and ethical considerations of different research fields, such as, for example, HC-based CS, by including functionality to keep ethical guidelines up-to-date. Moreover, it improves the review process for the main user groups (trustees and project representatives) by building on a HC approach to semi-automa-tically evaluate projects based on a rule engine and the adaptive ethical guidelines (and final human review). This not only makes the review more suitable as such, but also more efficient for project representatives who do not have to spend time with replying to questions that are not relevant or outright redundant. Moreover, the automated evaluation of research project before IRB experts take over allows them to focus on new and unexplored ethical issues that might come up in new research projects and do not loose time with administrative or formal aspects and well-known ethical issues, since these have already been addressed by the automated system. By keeping humans in-the-loop of the automated evaluation while at the same time continuously refining the underlying rule engine, the platform keeps adapting and improving. After the theoretical introduction and description of the platform the following chapter now turns to testing and evaluating this approach.

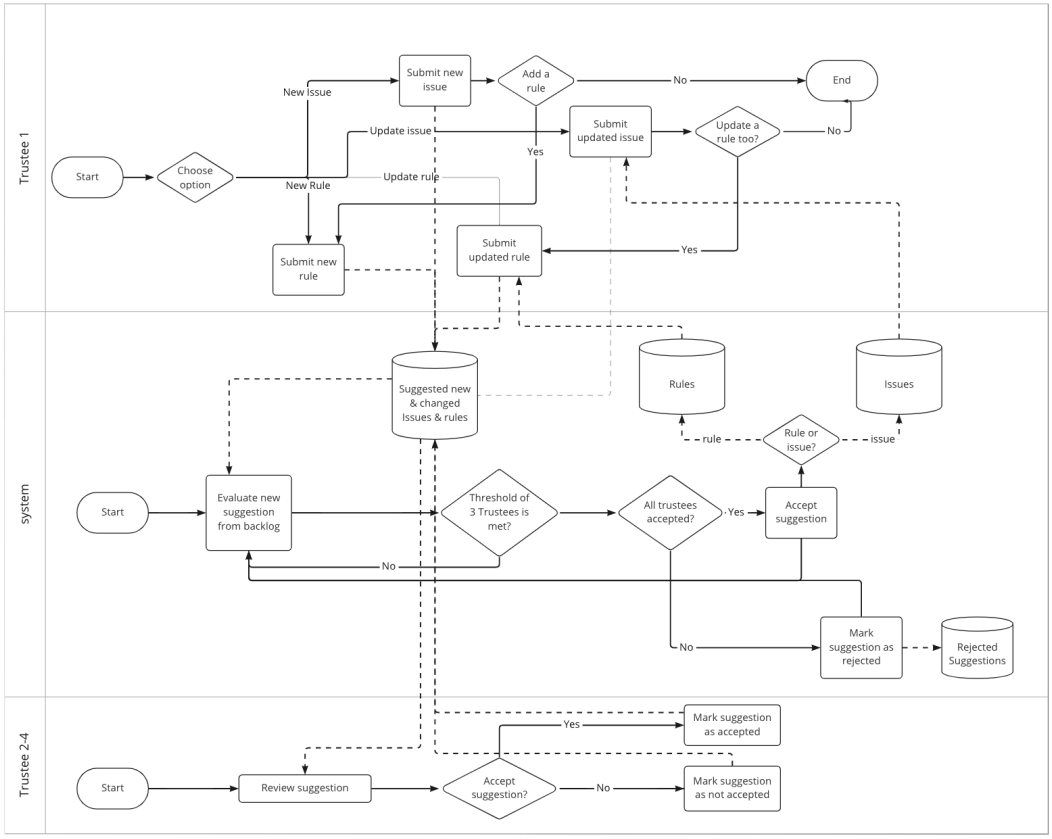


Figure 5.4: Process diagram for updating issues and rules

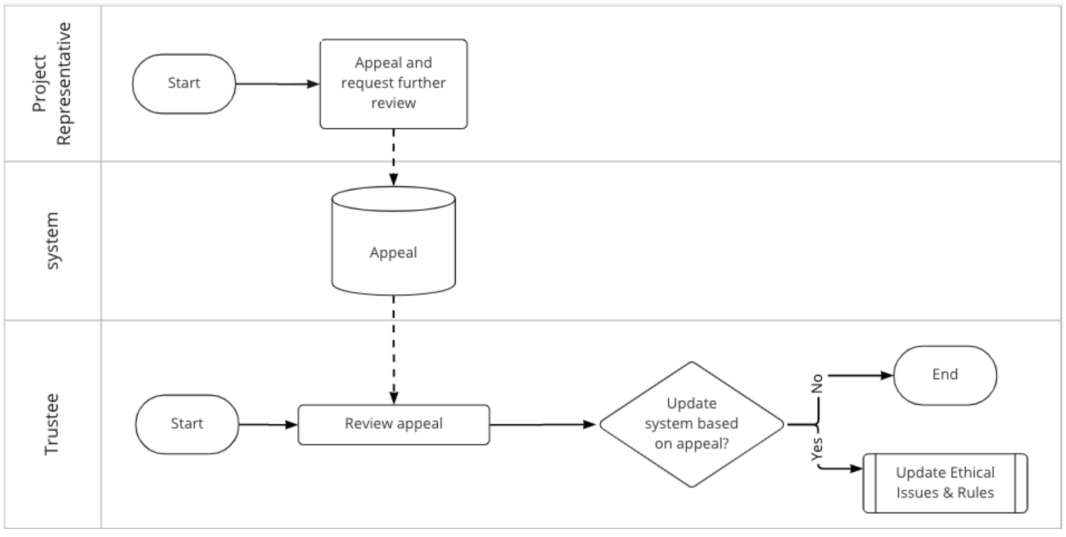


Figure 5.5: Process diagram for appealing evaluation decisions

CHAPTER 6

Validation of the new CAER system

6.1 Overview

This chapter describes the empirical validation of the envisioned new CAER system and its design. Since ethical review itself forms a complex process and the CAER platform addresses various aspects if its improvement, it is beyond the scope of this work to implement, test and evaluate a complete functional prototype of CAER. The aim of this work is to focus on the core features of CAER, which introduce new processes and functionalities to the ethical review process as described in chapter 5. More specifically, the following aspects are evaluated and validated in this chapter: The automated and dynamic evaluation of research projects and the maintenance and dynamic adaption of ethical guidelines by trustees.

To validate these main features, the following approach has been taken. Two proof-of-concept prototypes have been developed using *Airtable* (see chapter 2). These are tested and evaluated with a total of 12 test users as well as 2 IRB experts. In order to ensure comparable results of the test runs, concrete test scenarios have been developed which function as example research projects to be used during testing.
To validate that the CAER system more generally meets the application domain's needs and actually introduces what would be seen as improvements to ethical review in the U.S., qualitative interviews with two IRB experts with long-term experience in this area have been conducted, which allow to better grasp the bigger picture.
In the following, two possible test cases are presented, followed by a detailed description of interview questions, survey design and prototype development. Then, the sample of test persons is briefly described, before the results for the two prototypes are discussed separately. Finally, a summary of the validation results is given and discussed.

6.2 Case studies for validation

To evaluate the new CAER system and to ensure comparability of results, four possible case studies were designed, two of which are outlined in the following.[1] The selected test cases build in part on my own experience at the Human Computation Institute and present real existing or currently-under-development HC-based CS projects, which have been slightly modified for application in this work. The modifications are restricted to additionally added details to facilitate the evaluation process for test users. For example, even if such details were not fully defined in the real study, information on how participants are to be acknowledged in publications and how research results would be shared with them have been added to their descriptions here. The real project of the human-AI partnership test case has already undergone IRB review. This allows to include experiences with the current ethical review process by including questions from the original IRB review form in the design of the ethical evaluation form proof-of-concept. Moreover, designing example case studies from the context of the Human Computation Institute had the advantage that details on the case studies, which are not always publicly available, could be incorporated for evaluation purposes.

The validation sample size of 14 test users was selected as to not exceed the feasible scope of this master thesis, since the validation procedure itself was time intensive and further supplemented with qualitative analysis of the conversations with test users. While a larger validation regime may yield further insights, I believe the user testing conducted as part of this work is certainly sufficient to gain significant insights into the potential use and helpfulness of the CAER system. Of the 14 test users, the two IRB experts provided feedback on the prototypes without going through the full testing procedure, and two other test users evaluated the prototypes following the testing procedure but with their own HC-based CS projects rather than the example case studies presented here. See section 6.3.4 for more detail on the group of test users. In total, 5 test users were presented with case study 1 6.2.1 and 5 test users with case study 2 6.2.2, allowing a fair comparison of results.

Except for minor corrections, the descriptions of case studies presented below are reproduced verbatim to the descriptions that test users were given. That is to say: they are to be read as brief descriptions of research studies for which the test user will subsequently seek ethical approval.

6.2.1 Example case study 1: Human-AI partnership experiment in *Stall Catchers*

This case study is derived from a collaborative research study between the Human Computation Institute and Microsoft Research [133]. *Stall Catchers* is an online CS project, developed by the Human Computation Institute, in which participants contribute to Alzheimer's disease research conducted at a biomedical engineering laboratory at Cornell University by analyzing research data presented in a game format. Participants are presented with short movie sequences depicting blood vessels in the brains of mice and are asked to decide whether a selected blood vessel is flowing or stalled. Combining humans and machines to solve this analysis problem has proven to be very effective for *Stall Catchers* [134]. With the goal of better understanding how *Stall Catchers* can be made even more effective for solving the scientific problem to which it is dedicated and to make it more engaging for users, scenarios examining player performance in response to modification of user experience have been developed. To conduct an experiment, a sandboxed version of *Stall Catchers* has been

[1] The other two case studies were not ultimately used during validation and can be found in the appendix A.2 for future work and testing.

deployed in a separate environment so as not to influence actual research data. The aim of the experiment is to find out how the involvement of AI-bots assisting human participants in analyzing data influences the performance of humans and human-AI partnerships.

The experiment is designed as follows: Participants must first agree to the Terms and Privacy policy of *Stall Catchers* to register and then must actively agree to participate in the experiment. They confirm that they have read the consent form by activating the checkbox "I agree to participate" and indicating their date of birth (which is not stored but used only for restricting access for minors without parental consent and triggering the parental consent process) and clicking on the "begin" button. For individuals under 18 to participate in the experiment, parental or legal representative's consent via email is required.

Participants perform the same task as in the main *Stall Catchers* game, with the difference that they do not gain points and are asked to answer additional questions. Furthermore, an AI-assistance algorithm represented as a robot icon on the user interface points to one of the response options indicating its own predicted response. Participants can then make their choice accepting the AI recommendation or rejecting it. There are 4 different stages of the experiment. each of which consists of a certain number of movies to analyze. In each stage, different settings are used, and the sensitivity of the AI system is tuned to analyze the settings' impact on human and human-AI performance, as well as on how participants incorporate the suggestions. After each stage participants are asked to answer questions about their perceived own performance and the performance of the AI-assistant. After the last stage additional questions about their experience with the AI-assistant were included. In total, participation in the experiment should not exceed one hour and participants are free to take a short break in-between or to opt-out at any time. The first time participants contribute to the experiment, before the first stage, they are presented with an instruction page describing the experiment.

Publications resulting from this experiment will include *Stall Catchers* participants as co-authors and link to the list of usernames who contributed to the experiment. If participants do not want to be included in this list, they can opt-out via email. Moreover, aggregate results will be published in newsletter and online blogs and a link to them shared in the forum and in-game chat of *Stall Catchers* to make sure participants will be able to access the results. The AI model will not be publicly shared.

An open and public call for participation will be shared via the *Stall Catchers* platform. To participate, a computer or mobile device with internet connection is necessary. The experiment will be accessed via a web browser. The experiment will not pose any risks to participants beyond any normal risks associated with playing an online game. There is no medical research being conducted on the participants. The experiment only collects the email address as personally identifiable information of participants which will be stored on secure database servers. Server-to-server communication is encrypted with RSA.

6.2.2 Example case study 2: *CrowdMeter*

The third example case study is not an online game compared to the first two examples. In 2020 the Human Computation Institute started working on a response to the COVID-19 outbreak to allow people to actively avoid crowded spaces and thereby contributing to lowering transmission of viruses in society. *CrowdMeter* is a navigation app for mobile devices with the goal to reduce risk of transmission of COVID-19 in local areas [135]. Based on the assumption that crowded spaces increase the risk of transmitting viruses during an outbreak, *CrowdMeter* allows individuals to make decisions about where to buy groceries for example and helps to choose alternative destinations that is less crowded to minimize COVID-19 exposure risks [136]. *CrowdMeter* presents sets of similar destinations and for each destination is displays the respective risk of transmission (see figure 6.1). Additionally, it also displays times and predicted crowdedness at a certain time. This way, users can

set their expected arrival time and see how crowded it will be by then. The color green indicates no or little crowding, the color yellow describes that it is getting more crowded, and red means very crowded [137].

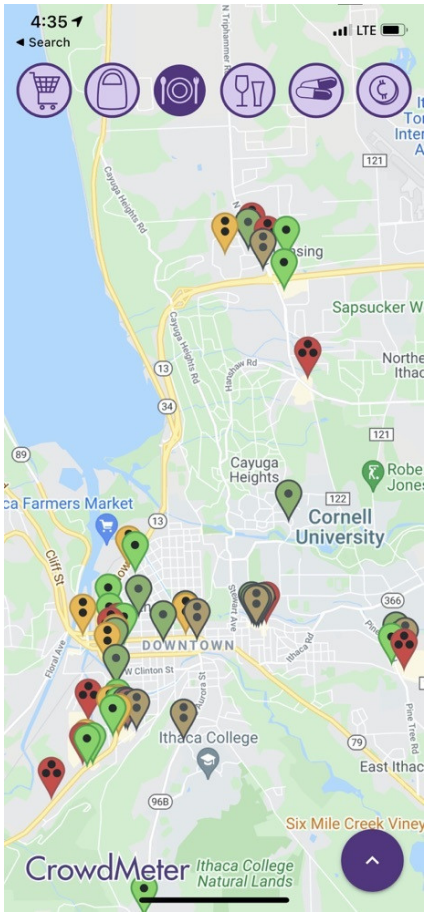


Figure 6.1: The *Crowdmeter* App ©Human Computation Institute 2022

CrowdMeter's algorithms are informed by previously conducted epidemiological and behavioral agent-based simulations [136]. These simulations showed that exposure rates could already be remarkably reduced if only 10 percent of the population of Ithaca, NY, the city where *CrowdMeter* was first launched, would use the app. With the support of the ASPEN technology Hub a collaboration of the Human Computation Institute together with researchers from Harvard University, ETH Zurich, Microsoft Research and Cornell's College of Veterinary Medicine conducted the initial research on *CrowdMeter*'s potential influence and to evaluate safety and plausibility of *CrowdMeter*'s idea [136]. The app was subsequently developed by the Human Computation Institute. Microsoft Research assisted with the development of the algorithms predicting future crowdedness.

With the release of the *CrowdMeter* app in Ithaca, NY, the goal is to collect feedback on user experience, to test the app's predictions and to better understand the impact of *CrowdMeter* on exposure to viruses in the real-world [137]. Except for the results of the simulations there do not exist any results on the latter.

CrowdMeter is available on iOS and Android. Participation is voluntary and participants can opt out at any time. Participants have to agree to the terms and conditions by marking a checkbox to use the app. They are informed about how their data will be used [135].

The use of *CrowdMeter* will not pose any personal risks to participants beyond any normal risks associated with using mobile apps and users decide for themselves where they prefer to go. However, it is not clear what would happen if all inhabitants of a city would use the app. It is for example conceivable that C*rowdMeter*'s approach could backfire because everybody would choose the same alternative destination. There is no medical research being conducted on the participants. The collected data is anonymized including usage data (such as location data and times and user feedback) and stored on secure database servers [135]. If participants decide to opt our, they can request to have their data removed.

6.3 Validation Procedure

6.3.1 Prototype design for first workflow: automated evaluation

To evaluate core aspects of the new ethical review process, a prototype for the automated evaluation of HC-based CS projects has been implemented with *Airtable* (see 6.2 and 6.3). This prototype implements one part of the envisioned collaborative and ethical review platform, namely the automated evaluation of projects. This is a key aspect of CAER that is particularly meant to improve the adaptivity and efficiency of ethical review. As has been discussed in chapter 2, current smart form builders allow the usage of conditional logic to only present questions relevant to a specific project dependent on the answers provided by a user. However, it is not possible to implement prediction based on previously evaluated projects and the project description into existing solutions which would additionally improve the adaptivity and efficiency of the ethical review. Hence, the presented prototype does not fully represent the envisioned functionalities, but it can nevertheless be used for the evaluation of the new approach.

For the prototype, a "smart" form ("dynamic form" in the following) was implemented with questions based on IRB submission forms. In total, 25 questions have been included, of which not all had to be answered in the dynamic form, as some "rules" make certain questions redundant under some circumstances (the included questions can be found in the appendix A.3). Dependent on the example test case applied, only those questions that were relevant to the test case therefore had to be answered in the testing procedure.
Since the aim is to evaluate the functionality of the prototype it is not strictly necessary to include *all* questions usually presented in IRB forms. Moreover, while project representatives in CAER would first have to provide information on the to-be-reviewed project including for example project objectives, background and research questions similarly to current IRB processes, this has been neglected in this prototype for testing purposes. Therefore, it has to be noted that the results of the automated ethical evaluation conducted with the prototype do not present "real-world" outcomes.

To compare the new approach with the current IRB process, a second form has been created (also using *Airtable*), which includes the same questions as the dynamic form, but does not include the conditional logic used to show questions that are contingent on previous responses (see 6.4). Test users therefore had to answer all 25 questions in this form. Test users were asked to fill in both the "smart" dynamic form and the "classical" static form, enabling the comparison of the user's experience with both forms and evaluation of the value of the new approach (without the prediction functionality).

Test users with experience in IRB review and/or (HC-based) CS either as IRB experts or project representatives were asked to fill in the forms based on a project for which they have or aim to seek ethical review (as noted above, two test users decided to do so). Test users without experience with IRB review as project representatives received a description

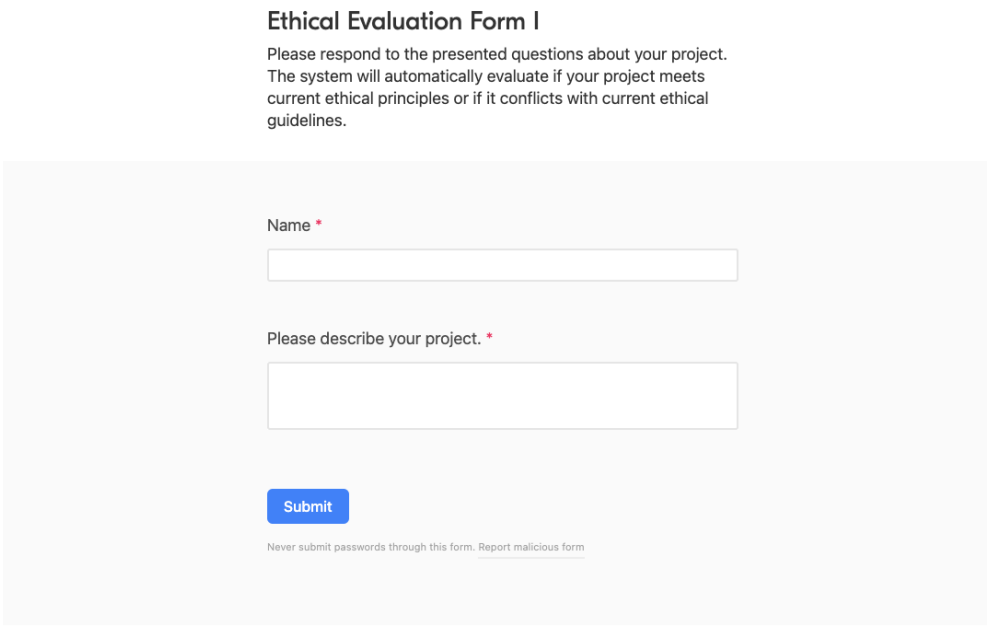


Figure 6.2: Excerpt of the dynamic ethical evaluation form prototype

of one of the example case studies and were asked to subsequently fill in both forms according to the provided information.

To give an example of the qualitative design of the case studies in relation to the questions posed in the forms: the user experience of the two forms when considering case study 1 (Human-AI partnership experiment in *Stall Catchers*) differed in their possible scope. In the "classical" IRB application, questions on physical training of the Principal Investigator had to be answered even though the research would be entirely carried out on the internet, i.e. even though these questions did not apply to this specific project, they had to be answered nevertheless. With the dynamic ethical evaluation form, project representatives would not have to answer questions focused on clinical research.
This prototype allows to further substantiate the approach and the results of this research and to help interview participants imagine the new collaborative and ethical review process.

6.3.2 Prototype design for second workflow: maintaining ethical guidelines

To make the review process more dynamic, efficient, and suitable to the field of HC-based CS, trustees should have the possibility to keep ethical guidelines up-to-date by making edits and changes to existing questions and rules that are processed by the rule engine, but also by suggesting new ethical issues and rules. This way, ethical issues that get introduced with new emergent technologies or the development of HC systems for example can be included in the evaluation to make sure ethical review covers all important aspects of research projects. To test this new feature of maintaining ethical guidelines, a proof-of-

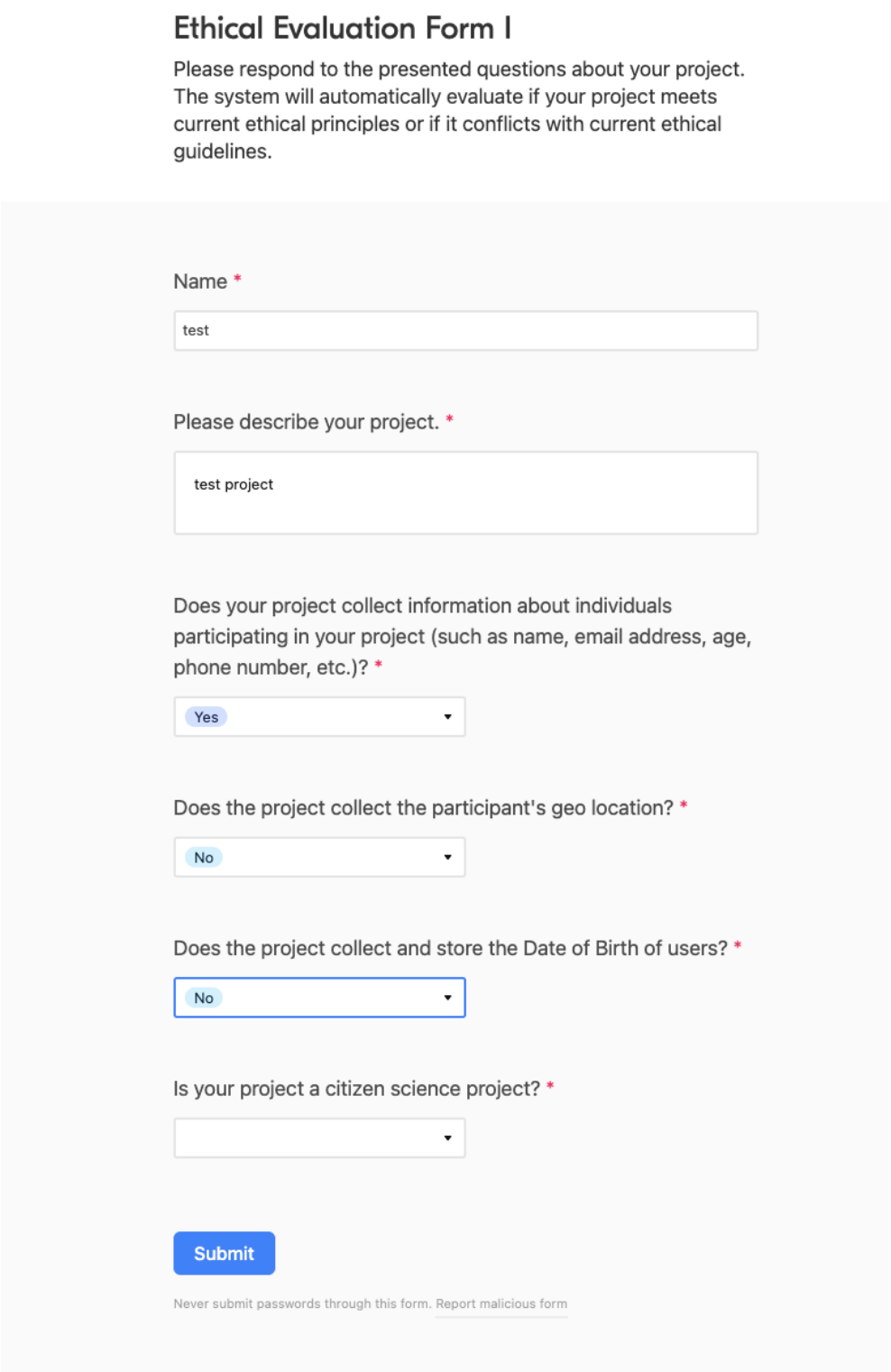


Figure 6.3: Excerpt of the dynamic ethical evaluation form prototype. The next question is presented according to the provided answer to the previous question.

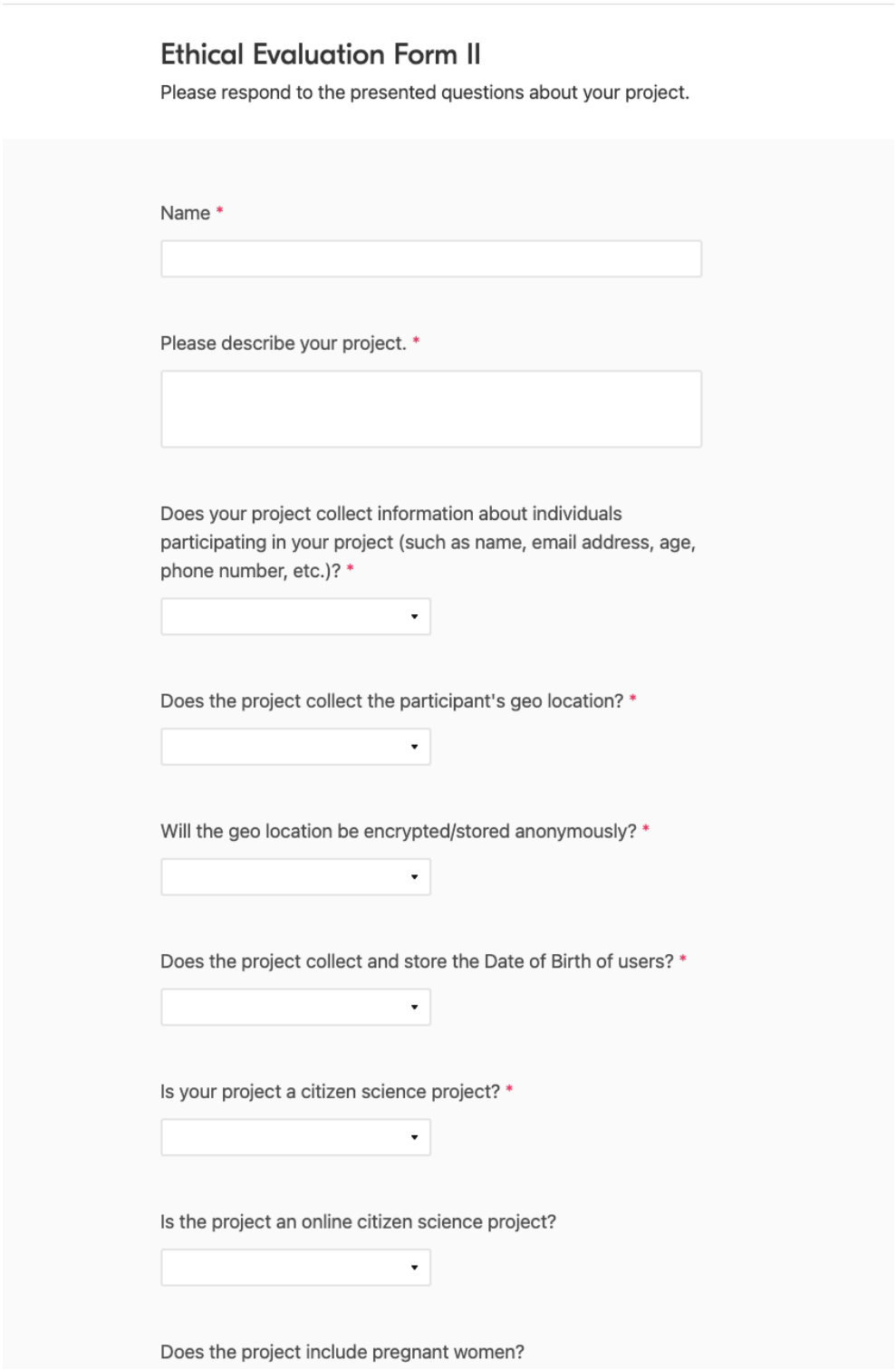


Figure 6.4: Excerpt of the static ("classical") ethical review form. All questions are presented at once and have to be answered independent of the to-be-reviewed project's domain.

concept prototype has been built, again using *Airtable* (see chapter 2). A notable difference to the prototype developed for the first workflow is that this prototype relies on *Airtable*'s Interfaces functionality, which enables the development of interactive user interfaces that enable users to both read data from and write data to an underlying relational database (as opposed to a form, which only writes a single row of data on submission) [138]. On the landing page of the interface, (see figure 6.5), users can choose whether they want to update an existing or create a new ethical question or to update an existing or create a new ethical rule.

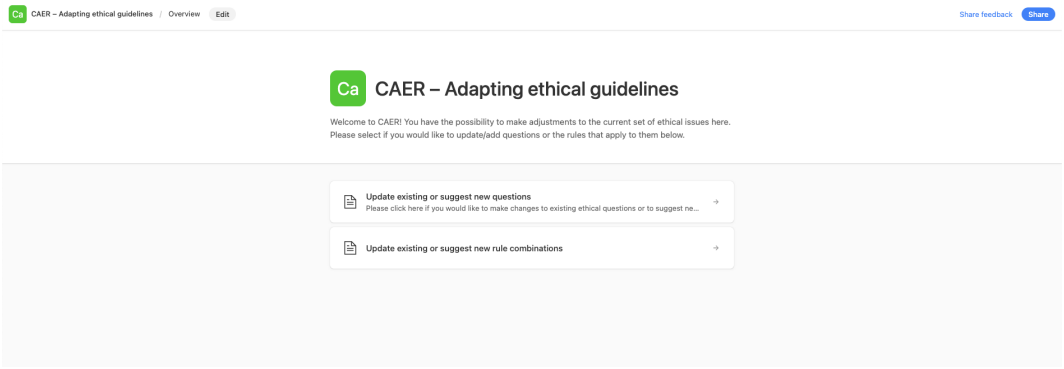


Figure 6.5: CAER: User interface of the startpage for maintaining ethical guidelines

If users choose to update existing or create new ethical questions, they are redirected to the interface for maintaining questions (see figure 6.6) where they are able to update existing ehtical questions, as well as create new ethical questions by clicking the corresponding button on the upper left side (which opens a form for adding a new question).

Existing ethical questions are displayed on the left side in a scrollable and searchable list element. To update existing questions, users can simply choose the corresponding question, change the description text in-line, and explain their changes in the text box below. Changes are stored automatically but not directly included into the existing questions. Instead, all changes are automatically tracked and the status of the updated question is changed to "pending for review" in the background. This would then trigger the mechanism to send out notifications to other trustees to review the changes before they get accepted into the current set of ethical questions. Note also that newly submitted and edited questions are not at all propagated to the form of the prototype for the first workflow - this is an important difference to how a real implementation of CAER would function. The two prototypes are totally isolated here, in a real implementation changes from this second workflow eventually have an impact on the questions presented to project representatives for the automated evaluation.

As mentioned above, upon clicking on the button on the upper left side, users are redirected to the user interface to create a new question (see figure 6.7).

Here, users can simply type their new question together with some explanation text and choose what answer options should be possible ("Yes", "No", "N/A"). In a real implementation of CAER, upon submitting the question, a mechanism to inform other trustees about the submission and request a review of the change would be triggered. Correspondingly, a new question would only included in the active set of ethical questions if enough trustees approve it, but as explained this dynamic does not apply to the prototype as implemented here.

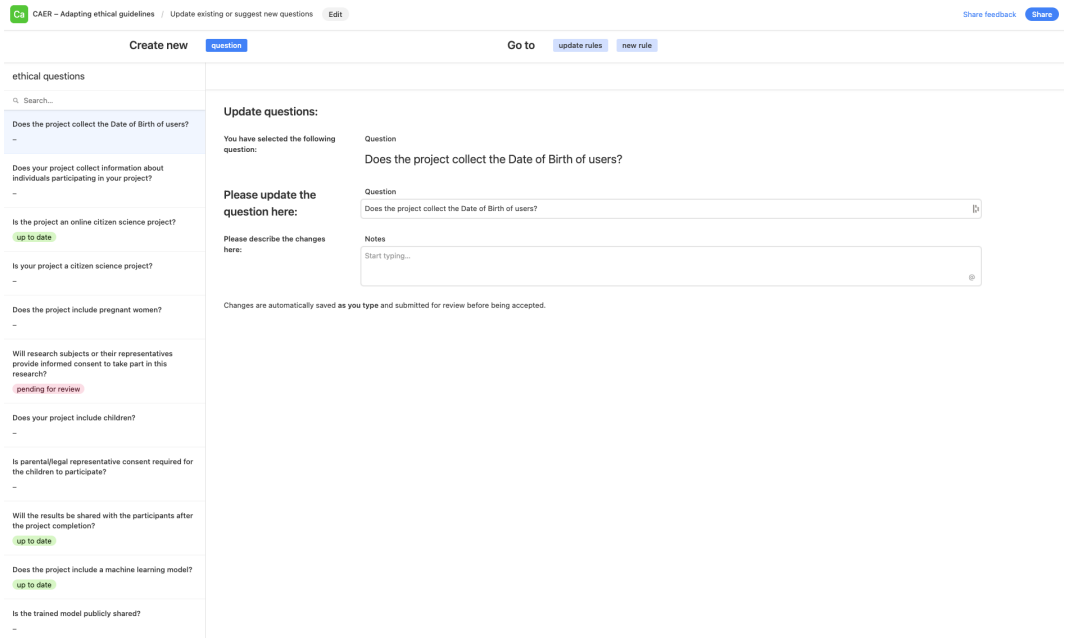


Figure 6.6: CAER: User interface to maintin ethical questions

After submission, users are led back to the user interface for maintaining ethical questions. If they want to create new or update existing rules, they can either go to the desired user interface by clicking on one of the buttons on the upper right side (see figure 6.6) or by going back to the main interface by navigating via clicking on the CAER symbol in the upper left corner.

On the user interface for maintaining rules (see figure 6.8), users can either update existing rules or go to the user interface to create a new rule by clicking on the button in the upper navigation bar, in much the same way as for question management. Again, current ethical rules are displayed in a scrollable and searchable list view on the left side. To update a rule, users can change the description of the rule in-line. Additionally, users can also change the status of a rule by for example "disabling" a rule. This means that the rule should no longer be considered by the rule engine. As with question management, a real implementation of CAER should store any changes as preliminary and trigger the corresponding review mechanisms.

To create a new rule, users can navigate to the dedicated user interface by clicking on the button on the upper navigation bar (see figure 6.9.

On the user interface to create new ethical rules, users first have the option to choose what kind of rule they want to create. They can choose whether the new rule is about 1) connecting an ethical question with a specific answer, 2) connecting two ethical questions, 3) connecting an ethical rule and a question, or 4) connecting two ethical rules. With these options, the system is very flexible and allows the creation of all possible combinations and for building very complex rules in an intuitive step-by-step way. If users choose the "about"-type of the rule, corresponding new options appear as shown in figure 6.10 with the example of connecting an ethical question with a specific answer.

Users can choose from the searchable list of ethical questions and answer options to connect these. Finally, they can select the option "leads to rejection", if the combination of

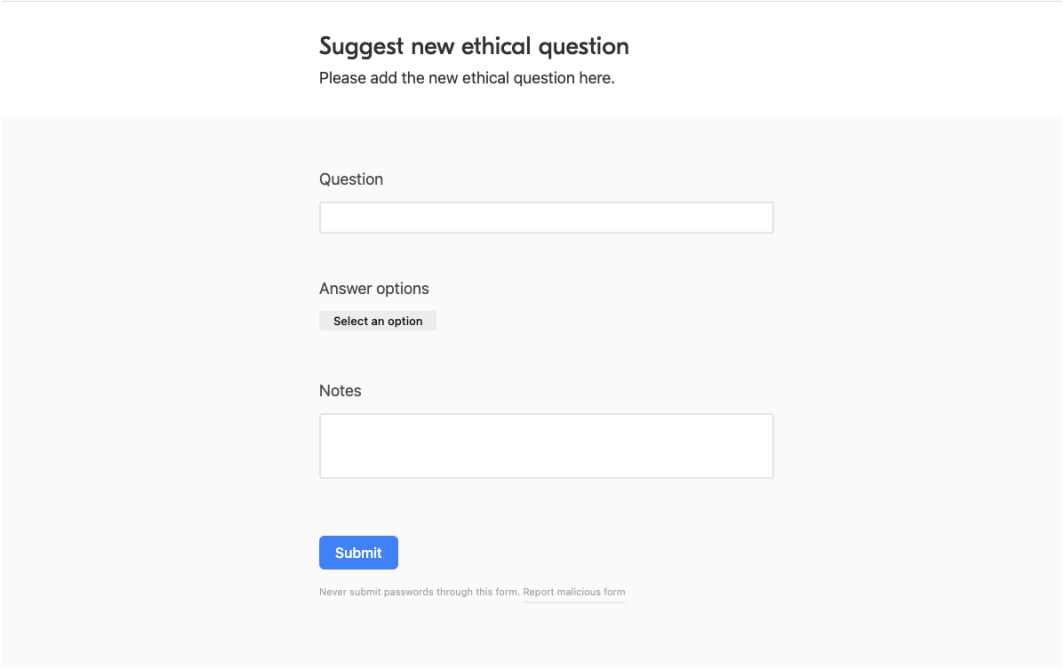


Figure 6.7: CAER: User interface to create new ethical questions

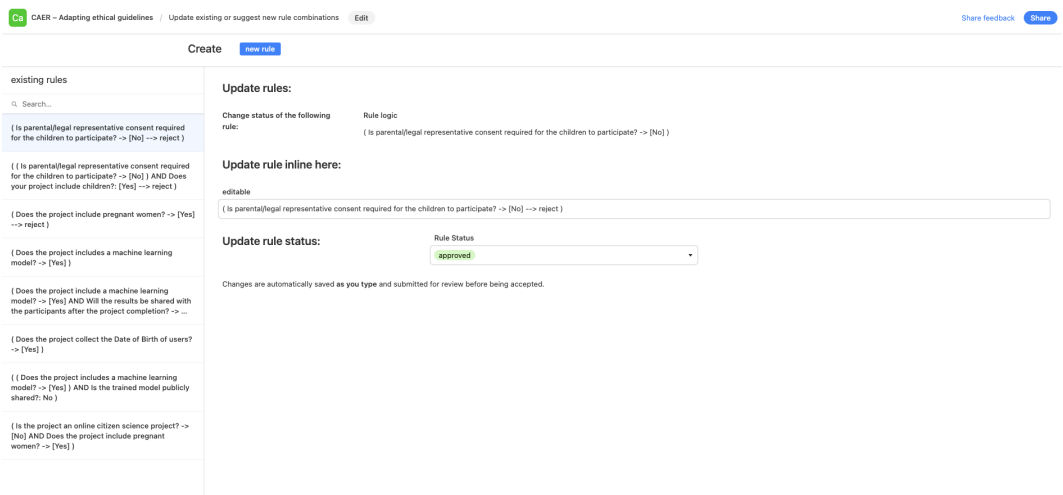


Figure 6.8: CAER: User interface to maintain and update ethical rules

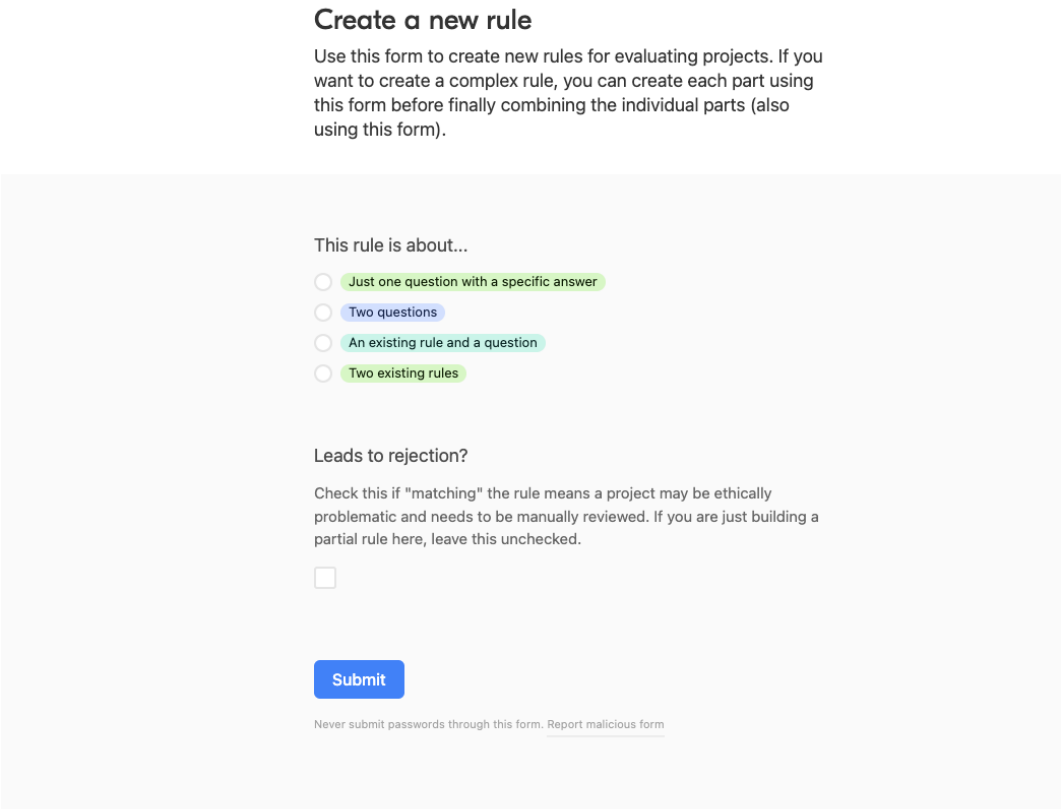


Figure 6.9: CAER: User interface to create ethical rules

the chosen ethical question and answer leads to the rejection of the project under review (see figure 6.11). The availability of this choice is required to enable the construction of more complex rules that combine multiple binary rules, where some included single binary rules by themselves should not lead to rejection.

The other forms of rules are created in a similar way. To connect questions or rules, users can additionally choose the rule type to be a logical “AND” or “OR” (see figure 6.12).

6.3.3 Survey design

To gain insights into user experience of the two prototypes but also a domain-specific understanding of the CAER system in general, semi-structured qualitative interviews were conducted. The advantage of semi-structured qualitative interviews is that they allow for the inclusion of follow-up questions and generally for asking open questions so that interview partners have the possibility to freely express their thoughts and experiences. Test persons tested the two prototypes one after the other, and for the first prototype, they also answered two different forms subsequently. After each form, they were asked to answer a few questions about their experience. However, after having completed both forms, they might change their perspective on the first form. For example, they might rate the experi-

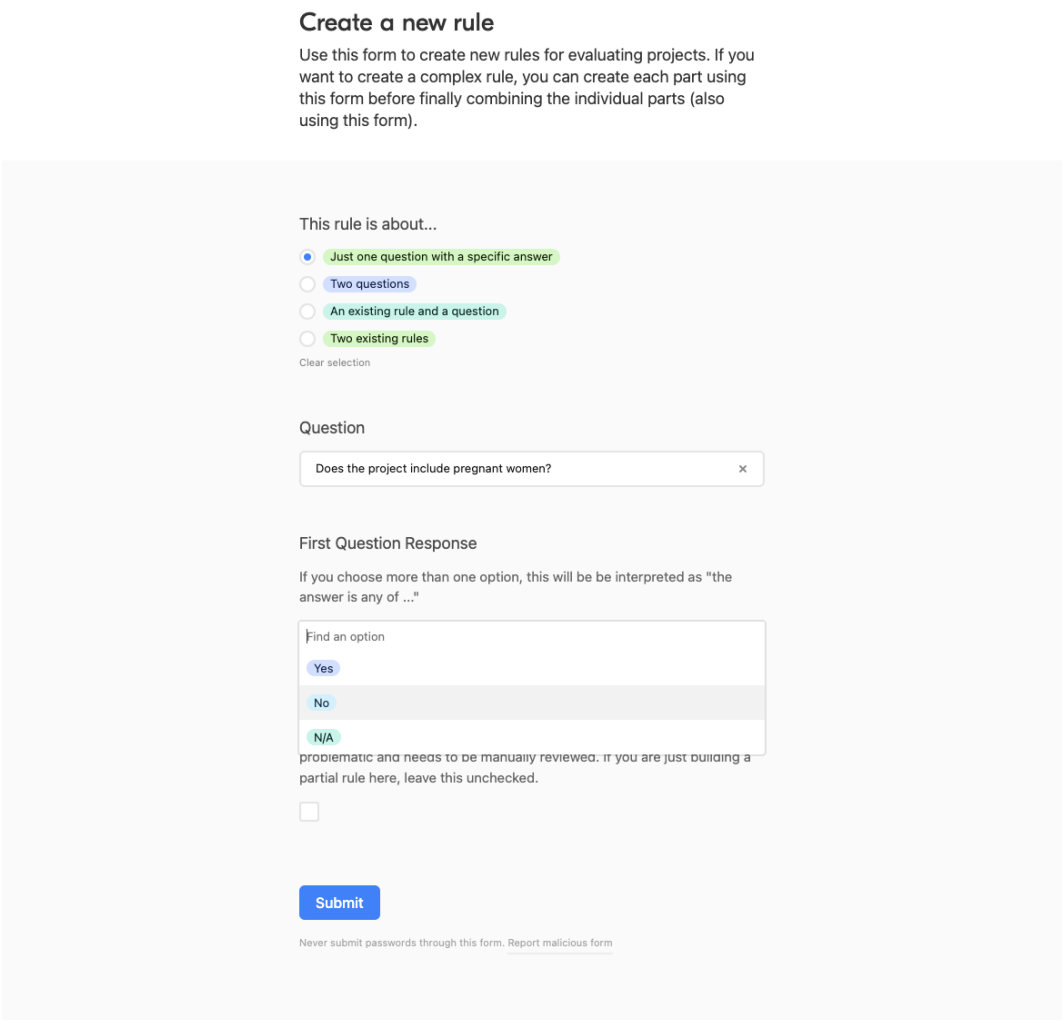


Figure 6.10: CAER: User interface to create ethical rules, choose from existing ethical questions

ence of the first form better before having completed the second form. An oral, qualitative interview allows to discuss this and to re-visit the responses without overriding the initial responses.

Two different questionnaires were developed: One for IRB experts which focuses more on the current ethical review process in the U.S. and how the CAER platform could help improve this, and another questionnaire for test users, which focuses more on the experience with the prototypes.[2].

The interviews were conducted via *Zoom*[3], audio-recorded and subsequently transcribed. The average length of the interviews was approximately 1 hour, with the shortest being 47

[2]Both questionnaires are attached to this thesis, see appendix A.4
[3]`https://zoom.us` [Accessed: Sept. 11, 2022].

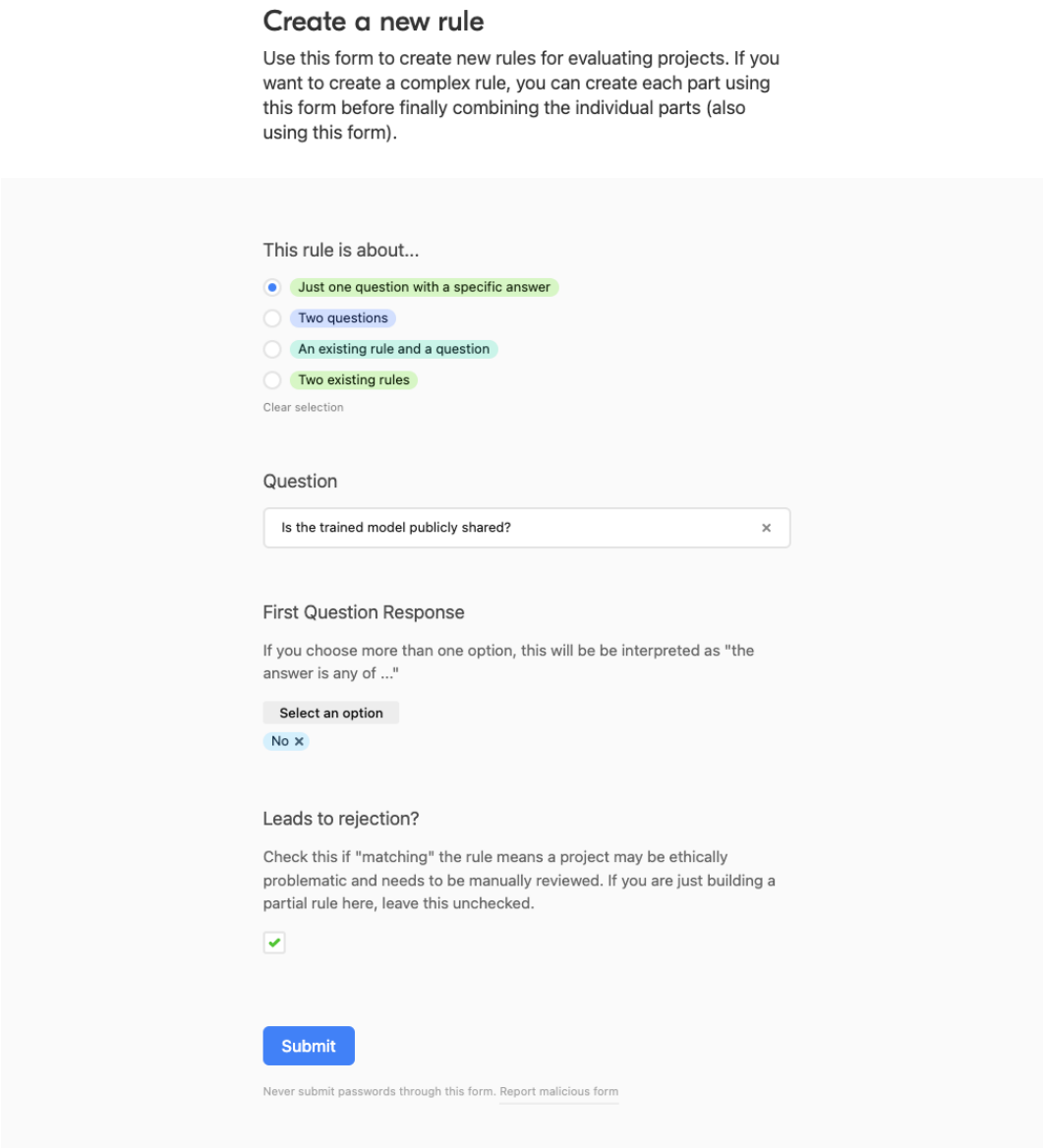


Figure 6.11: CAER: User interface to create ethical rules

minutes and the longest 128 minutes. Whenever possible, test users shared their screen during testing, which allowed me to follow closely how they interacted with the prototypes. In the case of the two IRB expert interviews, I shared my screen to present the prototypes. To ensure the privacy of the test users and IRB experts, all personal information has been anonymized in this work. In the following, quotes from the interviews have been slightly edited to improve readability but never substantially altered. One interview was conducted in German and translated by myself.

Besides the 10 semi-structured interviews and testing sessions, four additional remote

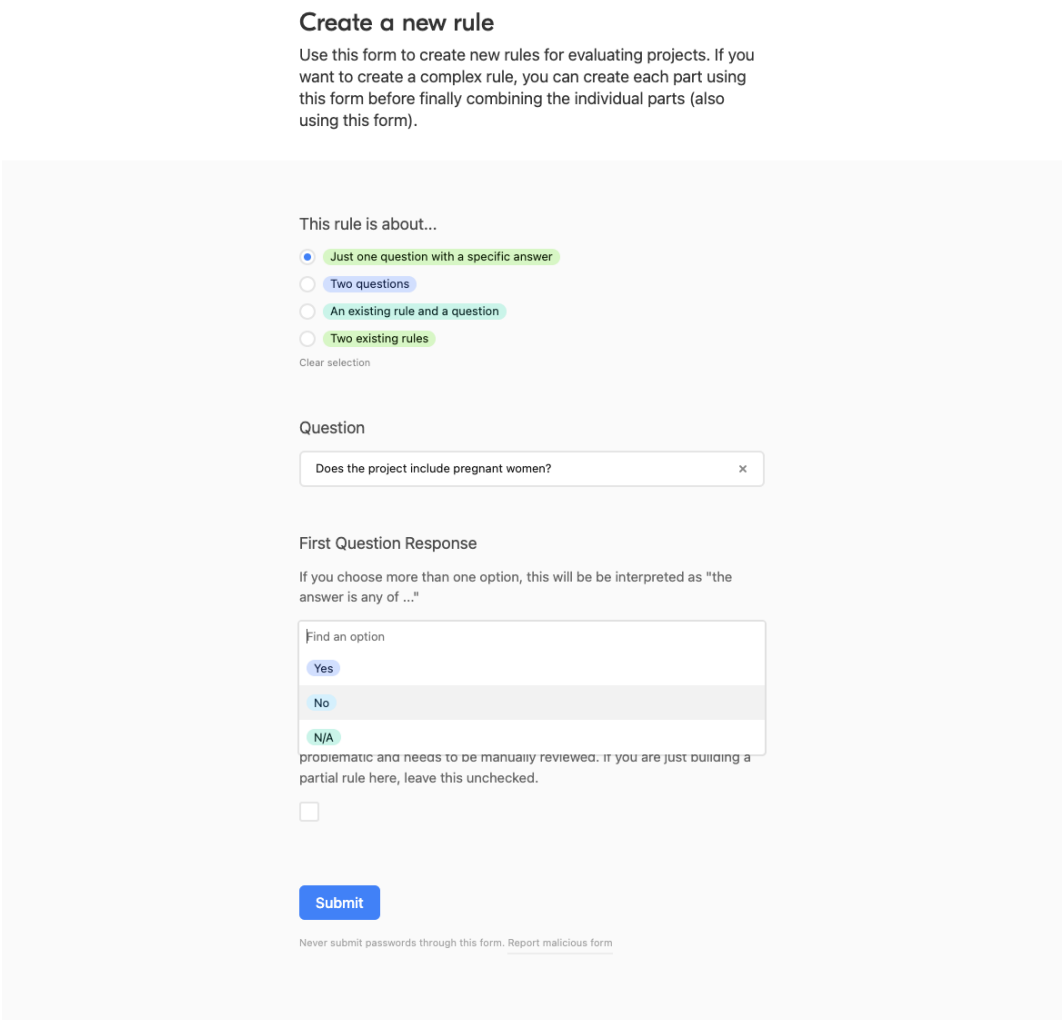


Figure 6.12: CAER: User interface to create ethical rules

testing sessions where test users were given written instructions and returned their responses in the form of written questionnaires were also conducted and are included in this work. This slight change in methodology had to be taken due to time constraints. Nevertheless, the design of the task was as close as possible to that of the synchronous testing sessions, ensuring the comparability of results from both groups.
In total, both example case studies presented in section 6.2 were applied five times each. Two participants used their own HC-based CS projects to evaluate the prototypes. The interviews with the IRB experts focused on questions on the current ethical review process and what could be improved and included discussion of the prototypes but no active testing of the prototypes was performed.
All study participants were asked to sign a consent form which informed them about the study and the handling of their data.

The test procedure started with a short introduction into the topic of this work. For the

first prototype, testing the adaptive automated evaluation of research projects, test users were first asked to read through the provided case study description. Then, they were presented with the first form, followed by questions on their experience, before moving on to the second form and subsequent questions about their experience with it. The order in which test users had to fill in both forms was alternated between participants, to control for biases towards either form due to starting or ending with a particular form.

For the second proof-of-concept prototype, testing the maintenance of ethical guidelines, after a short introductory note on the context of the prototype, test users were presented with six tasks to perform independently on the interactive prototype. The tasks were designed in such a way that, taken together, they covered the main application areas and practices to be performed to manage ethical guidelines by both creating new and updating existing ethical questions and guidelines. The tasks posed were as follows:

1. Suggest a new ethical question that asks if participants get acknowledged in project-related publications. Answer options should include "yes", "no", and "N/A".
2. Update the ethical question "Will subjects provide informed consent to take part in this research?" so that it includes not only subjects but also their representatives (representatives can also provide informed consent to take part in research).
3. Suggest a new rule where the question "Does the project collect the date of birth of users?" with the answer "yes" leads to the rejection of the project.
4. Suggest a new rule which combines the following existing rule with an existing question: rule: "(Does the project include a machine learning model? -> [Yes])" AND question "Is the trained model publicly shared?" with answer "no" LEADS TO rejection.
5. Update the following rule: "(Does the project include pregnant women? -> [Yes]); leads to rejection" by disabling the existing rule and suggesting a new rule which states that the question "Is the project an online citizen science project?" with answer "no" and the question "Does the project include pregnant women?" with answer "yes" leads to the rejection of the project.
6. Update a rule: The following existing rule contains a mistake because it rejects projects that should be accepted:."(Is your project a citizen science project? -> [Yes] AND Will the results of the project be shared with the participants? -> [Yes])". Please correct this in the following way so that a project will be rejected if it is a citizen science project and the results will not be shared with the participants.

After the completion of all tasks, test users were asked questions on their experience (see appendix A.4), focusing on the following aspects for each participant to ensure comparability of the feedback:

- Usability of the prototypes
- Comprehensibility of the navigation
- General feedback on positive and negative experiences
- Ideas for improvement
- Efficiency

Additionally, to evaluate the idea of the adaptive automated evaluation of research projects with prototype 1, test users were also asked to estimate the number of questions that were relevant to evaluate the ethics of their test research project. Finally, for prototype

2, test users were additionally asked, first, how comfortable they would feel to suggest new questions or rules and to update existing questions or rules in future, and second, if they would see any potential modifications to ethical guidelines this system could not address.

6.3.4 Sample of test users

To validate the envisioned CAER platform 14 interview participants and test users were recruited. Of these, 8 interview participants identified as female, 1 as female and non-binary and 5 as male. To cover user experiences from participants with various backgrounds and experiences, the participants' age ranged from 25 to 65 with the mean age of 41. Participants had various professional backgrounds and included 6 persons with a computer science background. All participants have a university degree, two hold doctorate degrees, therefore it can be said that a bias towards highly educated test users was present in the test sample, which should be taken into account when interpreting the results. Given that ethical review of research projects most often does involve individuals with some formal education, the sample of test persons can nevertheless be understood to be representative. Two of the interview participants have experience with HC-based CS as project designers and developers.

To evaluate the platform in comparison to the existing ethical review process in the U.S. two IRB experts with many years of experience in IRBs were invited (and agreed) to contribute to this research. IRB expert 1 is Certified IRB professional (CIP) and IRB operations expert and administrator with long-term IRB experience in academic institutions and hospitals, who currently works as a reviewer for a number of institutions. They have particular interest in "fixing IRBs, helping people discover what they can do more efficiently. That's what I really enjoy doing" (IRB expert 1, June 1, 2022: 5). Therefore, their input presents valuable insights for validating the approach and system design developed in this thesis. Both IRB experts have experience with founding IRBs. IRB expert 2 also has long-term experience with major IRBs in the United States and is currently a board chair of an IRB. The perspective of IRB expert 2 was not only helpful for this work due to their IRB experience, but also particularly due to their informatics background which makes them particularly qualified to evaluate how the envisioned CAER platform could add to existing software systems for ethical review.

It should be noted that it would have been desirable to include more interview partners and test persons with even more diverse backgrounds and experiences. However, more extensive testing and interview work was not feasible within the time frame and resource constraints of this work. While this could be improved upon in future work, the results presented here nevertheless serve as insightful first indicators for the utility and viability of the approach envisioned with CAER.

6.4 Results

In this section, the survey results for both prototypes are presented. While results are presented separately for each prototype in this section, a joint summary of both result sets follows in the following chapter in section 7.1.

6.4.1 Results of first prototype validation

This validation procedure involved two forms, one static form and one dynamic form, that were presented to participants in alternating order. First, feedback on each form is presented, before comparing the forms with regards to comprehensibility, efficiency, the perceived amount of relevant questions, and further general remarks by the test users.

Feedback on the content of the questions is neglected in the evaluation, since in practice, these are to be dynamically created and managed as part of the ongoing operation of the system and are not in that sense a fundamental part of the evaluation system itself.

6.4.1.1 Feedback on the static form

Regarding the overall impression and usability of the static form, test users reported that it was "quite user friendly" (TU6), "quite" (TU5) or "very usable" (TU3), and "pretty straightforward" (TU8, TU9). Several test users found it to be very "clear" (TU4, TU2, TU10) and well-structured. Thus, in general, the usability of form 2 was rated positively as "[u]ncomplicated, well structured and easy to handle." (TU2)

However, test users also described several issues they did not enjoy or appreciate. The first point of critique can be summarized with TU1's feedback: "Too many questions" (TU1; see also TU6). TU5 explained that they had the impression that they had to "answer the same question many times" at the end and questioned why some of the questions had been asked at all: "Because if it's not drug testing and it's not, why should I be asked if I am a physician? Why would it be relevant." (TU5) The impression of "too many questions" emerged from some questions not being relevant for the evaluated project: "I feel like I have to answer like a lot of stuff that's not even relevant that you can maybe rule out." (TU8) Test user 10, who evaluated an actual HC-based CS project they were currently developing, concluded that their research project "[f]its less into what maybe this survey [form 2] is geared towards, which [...] seems to be geared more towards clinical trials." (TU10).

Whereas the dynamic form would not pose a follow-up question on, for example, a project's policy on collection of users' geo-location once the project representative indicates that such geo-location would not be stored, under the static form, project representatives had to fill in all questions, regardless of their answers to previous questions. This was perceived to be confusing (see TU7). TU12 explained that this would make them doubt their own answers or question if they understood the previous question correctly (TU12). For TU11 it was also more difficult to "follow along" and more "distracting" compared to the dynamic form.

Suggestions to improve the static form, of which the main two will be briefly outlined here, therefore referenced the dynamic form and included the suggestion to reduce the number of questions by "[d]eriv[ing] next questions based on the previous answers." (TU1, see also TU1, TU3 and TU4) TU8 moreover suggested to include the possibility for project representatives to provide a written note for questions where they are not sure how to answer. This suggestion was particularly made for the static form due to the experience of uncertainty about how to answer questions that did not seem relevant to one's own project.

6.4.1.2 Feedback on the dynamic form

With the exception of test user TU9, all test users preferred the dynamic form over the static form, independent of which form they filled in first.[4] Responses included that it was "not too difficult" (TU5), simple, well-structured, clear and usable (TU1, TU2, TU6, TU7, TU11), "not overengineered" (TU1, see also TU4) and "[g]enerally easy to use and comprehensive" (TU2). Compared to the static form it was perceived as less overwhelming and faster (TU3, TU6, TU12) that could be filled in without experiencing any fatigue (TU10). TU7 summarized their experience as follows: "[I]n general I think it's very clear what to do and also that the questions, the next questions only appears if the last one is already answered. It really helps to stay focused on the current question. And yeah, I think it's very clear."

Moreover, all test users, including TU9, enjoyed the feedback feature of the dynamic form: "Okay, so I did like the feedback, the fact that once you put in a box and it's not in line or if, if indeed, your questions were in line with ethics." (TU9) Similarly, TU8 explained: "I like the idea of having [...] immediate feedback at the end because [...] in the last form, [...] you submit and then it is like okay what's happening now." (TU8) Receiving feedback on the project right away was experienced as more relevant, satisfying, and helpful in terms of knowing the direction of the research project (TU10).

Despite the overall positive feedback, two main concerns were raised by test users about the dynamic form. On the one hand, TU1 expressed they "got stuck at the final stage because my project was not aligned with current guidelines and I didn't know what option to choose." (TU1) This experience can be linked to the missing implementation of next steps, such as receiving more information on the decision or appealing or the ability to send a note to the human reviewer. These steps were not implemented to focus on the main features for comparison reasons (and also due to limitations of *Airtable* to implement them). On the other hand, due to the dynamic nature of the dynamic form, test users could not know in advance how many questions they would have to answer or how long it would take them to fill in the form. This was perceived as negative or mentioned as something that might be expressed to be a disadvantage by others (TU6).

This latter shortcoming directly links to a suggestion for improvement of the dynamic form made by test users: Including a progress bar that estimates the number of questions that remain to be answered (TU9, TU5). Since the idea of this form is to dynamically derive the next question based on the answer to the last question, such a progress bar could not be completely precise. Nevertheless, it could still be useful as a point of orientation and may be valuable to include in a full implementation of CAER. Even TU9, who preferred the static form, acknowledged that including such a progress bar "would be another way of doing it" (TU9) and would improve their impression of the dynamic form, since this was their main concern about the dynamic form anyway:

> "To be honest, I did like the [static form] a bit more because you had all the questions there. You had an idea of the number of questions and how long it would take you. And where you are within the, in the form. That's usually important for me to to understand how much longer I have to. How to stand it." (TU9)

Suggestions for improvement for this form focused on the progress bar, and some test users

[4] In one case (TU8), missing context information about the ethical review process in general led to misunderstandings which would then lead to first rating their experience with the dynamic form as worse than the static form. For example, TU8 did not initially understand what "appeal" or "more information on decision" referred to. After an explanation of these details, they subsequently updated their evaluation in favor of the dynamic form.

did not have any suggestions for improvements at all (for example, TU2 and TU3).

Finally, before turning to directly comparing results on both forms, a point raised by TU10 should be mentioned, since it points to another value of the dynamic form which had not been a particular focus of the design. TU10 explained that the dynamic evaluation procedure was "also helpful for [them] to start thinking about" (TU10) ethical issues that might come up in their project but which they had not considered so far. Filling in the dynamic form had an educational aspect, which was also mentioned by IRB expert 1 (see 6.4.1.4).

6.4.1.3 Comparative statistics

Having presented feedback on both forms individually, comparative statistics on the perceived efficiency and comprehensibility of the two forms as well as the estimated number of relevant questions that had to been answered allow to gain a better understanding of the affect of introducing dynamic forms.

In terms of the perceived efficiency of the forms, the average rating for static form, on a scale from 1 (worst) to 10 (best), was 8.1. In comparison, the average rating for the dynamic review form was slightly higher at 8.8.[5] TU8 summarized their experience as follows:

> "[The dynamic form] felt more efficient than the [static] form. Because [...] at one point where, like something happens that [...] makes the project not viable [...] under the [...] measurements you have. You get immediately this feedback and you don't have to bother with filling out ten more questions to get feedback because the ten questions that don't change anything at the end. So that's [...] more efficient than the [static] one." (TU8)

Similarly, the comprehensibility of the dynamic form was rated better than the static form. On a scale from 1 (worst) to 5 (best)[6], the dynamic review form) was rated with 4.5, whereas the static form was rated with 4.1.[7]

Finally, the biggest difference between the two forms was observed on the question of how many of the posed questions were considered to be relevant for the evaluated project. For the static form, test users estimated that 66% of the questions had been relevant. In comparison, the average for the dynamic form was 92%. From the test cases handed out to test users, in fact all questions contained in the "correctly filled" path of the dynamic form were in principle relevant for the evaluated projects. Therefore, it can be assumed that some test users were either not able to correctly estimate the number for the dynamic form due to lack of knowledge on the domain of the presented test case, or may have filled some answers incorrectly. 8 test users rated 100% of questions as relevant for the dynamic form.

The results show that overall, the dynamic review form was preferred over the static form by test users. It should be noted that the ethical review questionnaires simulated in testing were much shorter than real review forms tend to be. It can therefore be assumed that both forms, but especially the static form, were rated better than they would have been

[5] These ratings have been calculated over 9 test users. Test users who did not provide numbers for both forms were excluded from average calculation. Even though TU12 was omitted in this average since they did not provide an actual number, they rated both forms to be very efficient. However, the dynamic form was even "more efficient."

[6] The difference in scales used between the ratings for efficiency and comprehensibility (1 to 10 and 1 to 5, respectively), was an oversight in the survey design but should not be of great significance here.

[7] For TU5 and TU12 the numeric ratings on comprehensibility were added later upon further inquiry, since the scale had not been part of the survey from the beginning.

if test users would have had to answer all applicable ethical questions.
These results align with the feedback provided by the two IRB experts, which will be presented briefly in the following.

6.4.1.4 Expert feedback

In conversation with the IRB experts, I first explained the ideas behind CAER and then presented both prototypes and walked them through the individual forms. IRB expert 1 was very excited about the dynamic form and automated evaluation: "[T]his is brilliant [...] this is great. I really like it" (IRB expert 1, June 1, 2022: 73) and their IRB first impression of the dynamic form exactly hit its aim:

> "[W]hat you're doing is really important because you're taking, what you're automating are all the things that we shouldn't be thinking about and and actually are very rote and don't need a lot of time. And then I could actually think about ethical issues and maybe the system could suggest what issues are inherent in this particular study." (IRB expert 1, June 1, 2022: 27)

To IRB expert 2 the automated evaluation also "ma[de] a lot of sense" (IRB expert 2, June 7, 2022: 45).

Interestingly, their description of the current experience of project representatives with most IRB boards was very consistent with the above presented feedback on the dynamic form:

> "I worked with [different IRB] and they used the commercial system, and the forms are really anything but smart. [...] [A]nd investigators are actually basically asked to provide the rationale to the IRB. So, you know, we have the approval criteria. They basically ask to answer each one. They're asked a whole bunch of other things, and everybody is asked the same thing. And you can only imagine a researcher being asked to fill these things out, thinking, why am I even being asked? It doesn't make any sense. [...] I don't understand why that hasn't been done. It just seems like such a simple thing. Now, I think what you're also suggesting is that. When it's sent the submission to the [...] IRB reviewer, it would suggest some pathways like this is suitable for expedited review. This might be an exemption. You still have to make those determinations. I don't think there's anything wrong with that. I think that makes a lot of sense." (IRB expert 2, June 6, 2022: 29)

While this stresses the need to improve this process, IRB Expert 2 also emphasized the importance of human oversight over automated tools. In connection with this, IRB expert 1 particularly appreciated the "appeal" functionality included in the dynamic form (IRB expert 1, June 1, 2022: 64).

Beyond this positive feedback by both IRB experts, IRB expert 1 raised an important point (which also links to the results for the second prototype, presented below): IRB expert 1's suggestion was to use a different term for the "rules" underlying the dynamic review form, since "ethical rules" would sound too rigid and too strict for such a developing system and the nature of ethical guidelines: "Rules mean that you can't you either can or can't do it." (IRB expert 1, June 1, 2022: 65). The term "rules" had been introduced in reference to conditional logic that forms the basis of the computational system, but should be adapted to the application domain. A better term might be "directions" or "connections".

6.4.1.5 Suggestions for improvement for both forms

Before turning to the results on prototype 2, suggestions for improvement for both forms will be mentioned briefly. These can inform a possible full implementation of CAER, but can also contribute to a better understanding of problems with common forms currently used.

Several test users expressed the wish for more additional information on expert terms or why certain questions could be important for the ethics of a research project (TU4, TU8, TU11). This information could, for example, be presented in info boxes next to the individual questions.
Moreover, some test users felt that the answer options "yes", "no" and "N/A" were not descriptive enough or sufficient (TU8, TU5). Here, the prototype referred to common forms in ethical review but it could be considered to include additional answer options and to add a "note" section that would allow project representatives to include other information not covered or explain their answer (TU12, TU4).

According to TU11, they would moreover appreciate if questions would be grouped in categories. This indeed is a common feature of forms in ethical review which was not included in the prototype design. Since categories had not been introduced to either form, the lack of categories should not have affected the comparison between the two forms.
Other suggestions referred to the design of the user interface which could be improved by increasing font size etc. This unfortunately was not flexibly adaptable with *Airtable*.
Related to the last point, finally, two test users had difficulties with the user interface because it was not optimized for tablet computers (TU5, TU12), which relates to *Airtable*'s limitations. These last two points suggest that a customized user interface would be preferable.

6.4.2 Results for second prototype

In this section, results on testing the second proof-of-concept prototype for managing ethical guidelines are presented. First, general feedback on the usability and what the test users appreciated about the prototype are discussed, as well as general suggestions for improvements. After that, feedback on the clarity and comprehensibility of the navigation as well as the experienced efficiency of managing ethical guidelines are discussed. Finally, results on how comfortable the test users would feel to work with the platform with regards to managing ethical guidelines are presented.

It should be noted that some minor updates were made to the prototype after the first two test runs were completed. These updates did not affect the core functionalities to be tested, but slightly improved the overall user experience. The updates included a redirect link to the main "entry point" interface that was added in both user interfaces for suggesting new questions and rules. In order to make the interface for suggesting new rules more accessible on different devices (such as tablets or small computer screens), the display for choosing a rule basis was changed from a dropdown menu to a list view that shows the full text length, as *Airtable* doesn't allow changing the size of dropdown menus. This change was made after the first test user was not able to read the complete description. Lastly, an additional "suggest new rule" button in the update rule interface which used to be below rule description and status was deleted since it caused some confusion and the "suggest new rule" button in the header was sufficient for navigation.

6.4.2.1 Overall usability

Overall, all test users found the prototype to be easy to use and understand (e.g., TU1), clear, "manageable" (TU6), "logical" (TU7), "well-designed, clean" (TU10), "simple" (TU3), and "self-explanatory" (TU12). Test user 8 explained that "what I really like, just like the design of the combination of stuff like editing a rule or two rules or like, that's phrased pretty precise. [...] I saw it and I immediately knew what's like going on." (TU8)

Users were able to solve the presented tasks independently. Only in a few cases, some additional explanation of a task was necessary due to difficulties understanding the task description, missing context information or problems with internet connection or device used.
Some of the test users even described using the platform and solving the tasks as a fun activity (e.g. TU4, TU6, and TU7). For example, TU4 explained: "When setting up new rules in the empty mask, I got along fine. Selecting the relevant questions whose combination of specific answers should result in 'rejection' was fun."
For some test users, it initially took some time to familiarize themselves with the user interfaces but after that, they found it to be "easily understandable. Usage is apparent after 1-2 questions" (TU3), as test user TU3 explained.

6.4.2.2 Room for improvement and suggestions

Despite the overall good rating of the usability and experience, test users also reported some problems or issues they faced and suggested helpful improvements of which the main 6 will be discussed in the following.

First, not all test users found the prototype to be very intuitive (TU4). Several test users reported that the interfaces could be designed with more interactive user elements such as notes and different colors (e.g., TU1). Moreover, TU3 described the wish to include the "Go To" option on all pages so that users would not have to go back to the main menu. These suggestions relate to the current implementation in *Airtable* which has some limitations due to missing features and possibilities to customize the user interfaces. *Airtable*'s interfaces are also not optimized for mobile devices or tablet computers, which was also mentioned as negative experience by two test users who did testing on a tablet computer and had problems reading all button descriptions etc (TU5 and TU12).

Second, and also related to the implementation of the prototype in *Airtable*, test users complained about the experience for updating and creating information not being streamlined. Creating new questions and rules happened in a form-like interface whereas updating existing questions and rules is performed by changing text in-line. The in-line correction mode of questions and rules was confusing to some test users (TU5, TU4). After the first two test runs, the description to update the question by making changes in-line was enlarged and highlighted which helped in the following test runs. Moreover, whenever users switched to creating new questions or rules, a new tab was opened in the browser which would have to be actively closed again when the task was completed. Even though one test user (TU8) explained that the in-line editing of rules was also very efficient and intuitive, the experience could be made to be more streamlined in the final CAER platform. It would be feasible to include an additional text editor for updating rules and questions besides the streamlined experience for users who prefer the text-based solution. One suggestion for the streamlined solution by TU11 was to keep the form-style and allow users to drag and drop elements of rules to create new rules in addition to the drop-down menus

already used in the prototype.

Third, some test users expressed the wish for more explanatory text that would guide them through the individual steps for updating existing and creating new ethical questions and rules. This would also not be particularly difficult to accommodate in a possible implementation of the full CAER system.

Forth, TU12 described being hesitant about submitting new rules due to being afraid of making a mistake. The test user suggested including a preview of new rules, which could also be included for questions, so that users could see the newly created rule before confirming its submission (rather than submitting suggestions immediately).

Fifth, TU8 and TU10 suggested to include more response options in the question design, such as numeric responses. For the investigated prototype, the response options were chosen following existing ethical review forms. However, the idea of including additional answer options might be feasible, but would have to be evaluated carefully since this might increase the complexity of the system and the overall benefit is not yet well-understood.

Finally, and with an eye on the future, TU8 raised concern about the complexity of the system with increasing number of ethical questions and rules: here, better visual representation would be helpful. Moreover, it can be expected that this potential complexity would remain limited, since guidelines will not tend to grow indefinitely in their "depth" due to the nature of the problem tackled with the system.

6.4.2.3 Clarity of navigation

Regarding the clarity of the procedure in terms of how and where to perform the individual tasks, overall, test users found it to be clear and comprehensible. TU7, for example, said: "I think once you've found the menu thing, it was very clear because it's always there and you understand that you have to go back there to proceed." (TU7)

For TU8, the design was "obvious", but the interface could have included more visual elements:

> "That was quite obvious to me. [T]hey're basically [...] three buttons, which [...] are pretty [...] self speaking, [...] creating a question, go to [...] update rules or new rule. I would say [...] from an interface perspective, but I don't think [...] that's super relevant for your studies but from the interface perspective that could have [been] better. It took me like a second to understand, okay, this is [...] create a new question, this is [...] go to [...] update rules. But overall, there's three buttons to do an action. And on the left there's a search to [...] navigate what's there, which is quite obvious to me. I mean, the same as, if you click like update rule, you [...] just replace the search bar on the left with questions of rules which make sense." (TU8)

To improve clarity of navigation, test users mentioned the following issues and provided the following suggestions:

1. Include more visual elements to facilitate navigation.
2. Include extra views for the lists of current questions and rules
3. To one test user (TU11), it was not clear right away, how to proceed with the "this rule is about" options. However, none of the other test users mentioned problems here and also TU11 eventually managed to correctly use it independently.

All in all, these points could be improved upon in a full implementation CAER with reasonable effort. Features like the additional interfaces for displaying of current ethical questions and rules, are already outlined to be included in the platform design. This indeed describes an important feature, as the CAER platform should enable a transparent ethical review process and the displaying the current ethical guidelines visually for all platform users is a valuable tool to support this aim.

Furthermore, it may be particularly helpful to include an information box on the "create rule" interface that provides an explanation of how to create new rules to resolve typical problems encountered, as described by test user TU11 in the third issue above.

6.4.2.4 Comprehensibility of the navigation

On a scale of 1-5, the average rating for comprehensibility of the prototype was 3.5:

number of test users	comprehensibility rating
3	5
1	4.5
5	4
3	3

In explaining their rating of the prototype with score of 3, test users described difficulties with the user interface because they were using a tablet computer (TU5) for which the prototype could not be optimized, reported the wish for more clear visual elements (TU6) or explained that it took some time to get used to the platform (TU11, TU9). Once they had familiarized themselves with the platform, "[they were able to] actually navigate quite well" (TU11). Overall, test users found the prototype to be "pretty straightforward" (TU8).

6.4.2.5 Efficiency of managing ethical guidelines

In terms of how efficient test users experienced managing ethical guidelines, the average rating was 8.1:

number of test users	efficiency rating
1	10
1	9.5
1	9
2	8.5
2	8
3	7
1	5

The lower rating of 5 given by test user TU1 was unfortunately not further explained. Overall, TU1 reported being satisfied with the user experience of the prototype, and did not provide further feedback on specific features they disliked, nor suggestions for improvement.
Beyond this, explanations for not assigning the highest ratings included that "[it was s]ometimes unclear how to move between pages" (TU3) and that navigation was not always intuitive (TU5) (see comprehensibility of navigation above). Some explained they would have preferred a streamlined experience, but *Airtable* did not allow to handle question/rule

creation and adaption in the same way (TU6) and one test users stated that the "nature" of the task itself (creating rules and questions) would not be very efficient and hence the efficiency of the experience would be constrained, even though the design itself was very efficient (TU7).
High-rating test users explained that "[i]t was super fast" (TU12) and they "didn't feel [they] had to do [...] any unnecessary click" (TU8). One test user indicated that a focus on efficiency could have let to a trade-off for comprehensibility: "if that's [...] the most [...] Comprehensible way. I don't know, but it's definitely efficient." (TU11).

6.4.2.6 Future usage

All test users expressed that, after having completed the tasks, they would feel comfortable using the platform to create new and update existing ethical questions and rules in future. For example, TU9 explained: "Oh, I feel quite comfortable. Once you do find where things are, it's like any other platforms you have, you you get used to it and you are able to perform. I mean, since I could perform all these tasks. I would say I would be comfortable now." (TU9) Only one test user expressed concerns about performing more advanced operations in future (TU4). For a full CAER platform, it might be helpful to include features like the previously mentioned walk-through a tutorial in which users learn to navigate the platform and perform each individual task.

6.4.2.7 Expert feedback

Regarding the idea of dynamically updating and managing ethical guidelines, both IRB experts expressed that this would be a new path for ethical review which they had themselves never considered: "So I love this. I mean, I think the idea never occurred to me, which is the best kind of thing" (IRB expert 2, June 7, 2022: 55), explained IRB expert 2. Furthermore, they stressed that it would be "really needed" (IRB expert 1, June 1, 2022: 29–31) since ethical guidelines change over the years and have to be adapted to current regulations. IRB expert 1 provided the following example:

> "I think it's really needed because in the recent year or so, FDA has decided that pregnant women should be included. But now this whole history of pregnant women not being included and people are afraid to do it. So if we had a mechanism to gather data and say, see, this is you know, these are the instances where it's fine and there's no problem here, then we can start as you said, gather, you know, some sort of a baseline for why it should be changed, you know, but and also it would be useful to have to be able to help people understand why. So, you know, all right. So for years, pregnant women weren't involved. Okay? Now we're told they should be. So how do we get over that hump of, you know, people being afraid about it, seeing as a risk and actually discriminating against pregnant women? [...] So I think that's very much needed." (IRB expert 1, June 1, 2022: 29–31)

In a similar way, IRB expert 2 expressed their thoughts on prototype 2:

> "The idea of building another interface on top that would allow them to make suggestions to the forms, not only the investigators really, or the researchers or the board reviewers. Well, I should say not just the board of reviewers. I'd love to capture input this way from the people who are submitting this forms. [...] I think this is a really great idea because, you know, I make the forms and then, you know, two months later someone gets around to complaining about

> something and I go in and I never, you know, it's just one voice. And the idea that I could actually invite my IRB members to take a look at the rules and use their expertise to provide input. I think that's so cool. And I have to think about how I would do it because I'd love to do it, but I think it's a great idea." (IRB expert 2, June 7, 2022: 55-57)

Besides this positive feedback, the IRB experts mentioned other advantages that had not been actively considered as part of the CAER platform but which could be focused on in development of a full version of the platform.
For example, IRB expert 1 mentioned the educational value this platform could have, especially due to inviting community members and different trustees to discuss new issues and questions and dynamically adapt ethical guidelines (IRB expert 1, June 1, 2022: 33). Here, the feature of being able to leave explanatory notes would be important, so that trustees can explain why a question or rule should be included. This explanation should then also be shared with project representatives during the automated evaluation process by including information boxes on the respective questions (ibid.: 31).
Another aspect brought up by IRB expert 1 is the fact that reviewers could include discussions on questions and issues:

> "And then that in itself is becoming part of the discussion of what the ethical issues are and how people view them. So it, it, you know does educate, but it also allows the person who's doing the review to take all that into consideration. Maybe it's something I didn't think about before. I like it. I think you're on the right track." (IRB expert 1, June 1, 2022: 56)

Finally, IRB expert 2 mentioned another important detail, which is that even though some specific ethical issues (such as whether pregnant women are included in a research project) might not apply to a certain area of research, they still need to be considered in general if federal funding is involved. A possible approach to integrating such a requirement could be that such contextual questions be presented to a project representative the first time they use the platform, and if their research area is not affected by these specific aspects, the information could be stored and automatically be included the next time they apply for ethical review:

> "You know, you considered these requirements and they didn't apply, so you wouldn't have to do it with each study you saw. And I think in a way, that's what you're talking about, guiding into those categories. Which is a great idea that I should do. [...] So I think that's good." (IRB expert 2, June 7, 2022: 31)

After having presented the results for both prototypes individually, they are discussed comparatively and together with the implications of the results in the next chapter, followed by the analysis of the broader (societal) impact of the new approach to ethical review, the limitations of this thesis and future work.

CHAPTER 7

Discussion

7.1 Insights from prototype testing and expert interviews

While[1] both forms were perceived as user-friendly and comprehensible, the "classical" form was described to be tedious due to too many questions in general and particularly too many questions that were not relevant to the evaluated projects. Given that the prototype included only 25 questions (of approximately 50 questions project representatives have to answer in the "Initial Review Submission Form" by WCG IRB [139]), it is to be expected that this impression would only increase in "real-world" ethical review application settings.

Only one test user preferred the static form. The main argument against the dynamic form (form 1) for test user TU9 was the unpredictability of how long the evaluation would take. This point of critique is very comprehensible and should be considered in the development of the full technical prototype of CAER, for example by including a progress bar.

The dynamic form was moreover described to be less overwhelming, faster, and particularly valuable due to the feedback feature. Receiving feedback was perceived as very satisfactory and helpful for knowing the direction of the research project.

In terms of efficiency and comprehensibility, the dynamic form was rated more highly than the static form. Most important in terms of evaluating the value of the new dynamic evaluation process is the comparison of the results on the perceived efficiency of the two forms. Here, the very clear preference of the dynamic form shows that it presents an improvement to the still prevalent "classical" application process in which all questions, no matter how relevant they are for a specific domain of research, have to be answered. Moreover, as has been confirmed in the IRB expert interviews, this approach would also facilitate the reviewers (here: trustees) duties, since they could focus on those ethical issues that have not been addressed in other ethical review cases before and would not loose time on already discussed and clarified ethical issues.

[1] Many of the immediate items of feedback and suggestions provided by the test users and IRB experts have already been briefly discussed in a direct way in the previous chapter. Not all of these items are repeated and covered in depth here, but their significance should nevertheless not be understated.

Both IRB experts also stressed the importance of human judgement, particularly in relation to ethical review. In fact, CAER's HC-based approach guarantees that humans remain in control of the decisions the semi-automated evaluation ultimately arrives at, since they provide the questions and rules. Crucially, to make sure that no ethical issues are overlooked, the automatically created result is still reviewed by a human expert. Additionally, project representatives can get in contact with human reviewers if they are concerned about some particular issue, or if they disagree with the outcome of the evaluation of their project (appeal functionality). This was also acknowledged as important by IRB expert 1 (as stated in 6.4.1.4).

Despite the identified room for improvement, all test users were able to solve the presented tasks and described the prototype for the maintenance of ethical guidelines as usable and some even as fun. Regarding the comprehensibility of the navigation, the prototype was rated with 3,5 (out of 5). Most remarks could be resolved by focusing more on user-friendly user interface design for CAER. The possibilities to create clear and well-arranged user interfaces was restricted with *Airtable*. In terms of efficiency of managing ethical guidelines, the overall rating was 8,1 (out of 10), which shows that the approach to creating new ethical questions and rules as well as updating them is very promising and an actual path to keeping ethical guidelines up-to-date and adapt them to the needs of emerging research fields. This has further been confirmed in the interviews with the IRB experts.

Finally, the results of the evaluation of the two proof-of-concept prototypes show that the newly designed features do not require specific expertise in ethical review. All test users were able to perform tasks and navigate the presented prototypes. This is very important for the improvement of the ethical review process since the CAER system should be accessible not only to reviewers and researchers but also to research participants and citizen scientists to make the process more transparent and equal.

With regard to the goals given at the beginning of this work, to investigate the *suitability, adaptivity, equity* and *efficiency* or CAER, it can be said that, though still subject to further research and testing, these goals have largely been met. The suitability of the approach outlined here was directly confirmed through the conducted expert interviews and user tests of the prototypes. The adaptivity of CAER similarly was supported empirically, and to some extent can also be considered as self-evident in the design of CAER itself. In fairness, equity of a software system is difficult to properly evaluate before the tool is actually in operation, but here there were also indication in the expert interviews that CAER has potential in this regard. Finally, efficiency of the approach was tested directly in the user tests, and the results suggest that CAER's approach does indeed improve the efficiency of the ethical review process.

7.2 Broader (societal) impact of new approach to ethical review

The emergence of new research fields, such as HC-based CS, along with recent advancements in AI research in general (as discussed in this work, particularly in chapter 4 and the related work section 2), have made clear the importance of keeping ethical guidelines up-to-date and adapting them to the research domain of the to-be-evaluated research projects. Current ethical review, however, cannot easily (if at all) dynamically adapt to the needs and ethical issues of these emergent fields. Therefore, this thesis suggested a new dynamic and adaptive collaborative ethical review platform. The CAER platform could thereby not only introduce new ways of doing ethical review, but could furthermore enable and support a much-needed discussion on ethics for AI and other emergent research fields as has been described by IRB expert 1 (IRB expert 1, June 1, 2022: 24-25).

The new CAER platform could contribute to this discussion and would, at the same time, function as an educational tool for researchers (also called project representatives in this thesis), since they would learn about potential harms and ethical issues during the evaluation of their projects (see chapter 6). Moreover, since the CAER platform is to be transparent, all users of the platform could see and understand the current ethical guidelines at any given time.
Therefore, the new CAER platform could function as a hub for such discussions on, for example, AI ethics, which is accessible to everybody, not only IRB experts and researchers, but equally to participants. Actively inviting participants or potentially also human subjects, and letting them suggest new ethical questions and rules the platform, aims at lowering the barrier for them to contribute and to share their perspectives. Hence, this could help to reduce the power hierarchies described in chapter 4 and improve the discussion between IRB experts and community representatives by mediating the communication will via the platform. Nevertheless, including community representatives was described as one of the main challenges for changing the ethical review system by the two IRB experts. IRB expert 2 explained: "I think at least we need something that is able to listen to different groups. That's always a challenge because, you know, there are groups, there are lots of groups. [...] and when you you can't listen to everybody all the time. And I think that's a real challenge. I don't know." (IRB expert 2, June 7, 2022: 19) However, including community representatives is not a trivial task, IRB expert 2 goes on to explain:

> "Now, if you tried to have everybody represented on an IRB, you'd never get anything done. And typically, the people who would volunteer to be so representative would be the ones who were most stridently sort of in one direction. So I don't think representation is the answer. I do think that some of the things people are talking about in the United States outside the context of the IRB system. Community engagement. So reaching out to people, reaching out to communities when you're designing the research." (IRB expert 2, June 7, 2022: 19)

Additionally, new emergent fields of research such as AI and Big Data require a public debate, since they cannot be "policed" from one single institution, as IRB expert 2 argued (IRB expert 2, June 7, 2022: 21). While the CAER platform could contribute to this debate and open the discussion to more groups of people than before, the aspect of the engagement of community representatives also leads to the limitations of the prototypes and design of the CAER platform outlined in this work, which are discussed in the following.

7.3 Limitations

Due to the scope of this master thesis, it was not possible to describe all ultimately necessary functionalities of a "real-world" implementation of CAER in full detail. Therefore, the focus, also in the implementation of the proof-of-concepts prototypes, was placed on the core aspects of HC-based ethical review processes and workflows. However, this led to some limitations in the efforts undertaken in this work, such as the question of community representation not being discussed in enough detail. IRB expert 2 did reflect on the dilemma of this question during the interview:

> "Every research participant is different. Every value system is different. How you choose your community members. [...] [P]eople, lay people who choose to be involved in this are typically research advocates. You know, if anything, more than investigators want the research to go forward because they had a

> child who suffered a genetic problem or you. So [...] they're hardly representative. It's people in the general public don't think about this until they get sick. And then, you know, it's different. So that's really hard. I think the idea is spot on it, but how do you do it, is, is difficult." (IRB expert 2, June 7, 2022: 31)

Tackling this question would certainly need more attention and should be approached from many different angles, something that this work alone cannot accomplish.

The proof-of-concept prototypes implemented in this thesis also had some limitations. Although *Airtable* allowed for building the proof-of-concept prototypes with relative ease, it also entailed certain constraints in regards to the functionality and user experience of the prototypes. One major factor in this was the fact that the interfaces could not be customized very flexibly. The shortcomings of this were evident in the received user feedback at multiple points. Many test users reported on issues on the interface such as the size of buttons, the colors used, or arrangement of the interface. Though minor, these issues might have distracted or encumbered the test users with regards to the main aims of the tests, which was the evaluation of the overall functionality, efficiency and comprehensibility of the prototypes.
In *Airtable*, it was also not possible to fully connect the functionality of both prototypes to implement a joint system that allowed for semi-automatic evaluation of research projects and maintenance of ethical guidelines with a rule engine. Even though this was not visible to the test users, and the very same ethical questions and rules were included in both prototypes to create a uniform experience for the test users, logically, the two prototypes were completely separate from one another. Therefore, the platforms main functionalities could not truly be tested together. The behavior and usability of such a joint system should be investigated in more detail before, or as part of, a full implementation of CAER.

Finally, another limitation of this work might be presented by the sample of test users who participated in testing the prototypes. Most of the test users (10) did not have domain knowledge about HC or CS, and also did not have much (or any) experience with ethical review in the US. Partly, however, this missing domain knowledge could be compensated for via the IRB expert interviews and by including two test users with experience in the development of HC-based CS projects.
Test users without any experience received example research projects as test cases. It is not completely clear whether the information provided to them was always sufficient for them to have a thorough understanding of the processes and the domain, or if more information should have been included. The extent of this problem of course depends to a large degree on the individual test users and their background. At the same time, however, the use of example research projects allowed to compare the test runs, which would not have been possible on the same level, if all test users would have user their own real research project.

7.4 Future work

Future work in this line of research should explore several aspects of a new collaborative and adaptive ethical review process. To begin with, it would be advisable to develop and test a full technical prototype of the whole CAER platform, taking into account the features specified in this work that were not part of the prototypes presented here. Testing of such a system should include representative test users from all stakeholder groups, including CS participants. Subsequently, contingent on successful tests, a full development effort of a "production-grade" implementation should be estimated, along with further investigation of how this new approach to ethical review would be accepted by the domain and other

IRBs.
Moreover, additional research should be conducted on the involvement of community members and participants in the ethical review process.
More closely related to the individual features of CAER, the automated ethical evaluation could be further improved by an additional statistical component that tracks how often project representatives from specific research domains would choose specific answers to certain questions, leading to the rejection of the project. This knowledge could be used to make the automated evaluation even more efficient by optimizing the order of questions based on the statistical distribution of question-answer rejection patterns observed.
Related to this point, the automated evaluation could also potentially further be optimized by leveraging Natural Language Processing (NLP) in the analysis of the project descriptions provided by the project representatives. For example, it might be possible to speed up the ethical evaluation by applying automated question answering using large language models, which perform exceedingly well at zero-shot question answering tasks. However, such a step would have to be investigated very carefully, since this may partly be perceived to remove humans from a key part of the loop, so it would have to be ensured that nevertheless, no questions are missed or answered incorrectly. The feasibility of this would also strongly depend on the quality of the project description, introducing a new problem to tackle, namely how to educate project representatives to write meaningful and complete project descriptions.

Another aspect that is important to analyze in more detail may be the development of a suitable consensus mechanism for reviewing newly suggested or updated ethical questions and rules. While it may be simple enough to initially set the required number of trustees to review suggestions to some arbitrary number, it should be analyzed what number of trustees actually leads to a sufficiently dynamic but still robust system, such that decisions to change guidelines are practically feasible but also not haphazardly applied. This means that the decisions to accept or reject suggestions should represent the majority of trustees without creating too much work.

Finally, an aspect which has only been touched upon briefly so far in this work, but which should play an important role in future work, is the exploration of features centered around saving the automatic project evaluations to build an archive of evaluated projects, and to leverage this archive in the broader ethical review process. This may also increase transparency and could contribute to the education about ethical review. It could also contribute to more comparable review processes and decisions across different IRBs, since reviewers could consult this public repository of previous reviews on similar issues before coming to a decision. Such a course would begin to address the problem described by [40], namely that IRBs today come to very different decisions because "each board has a unique history of cases, members, and struggles" [40, p. 165] which inform these decisions. Privacy, data security as well as intellectual property concerns are a complicating factor in this but should not in principle prevent such an undertaking from being feasible.

These suggestions for future work suggest that this work serves as a basis for future research on computational approaches to improving the ethical review process and further discussion on this process itself. As the expert interviews and the literature review have made clear, there is a need for a different approach. This work lays the foundation for a shift from the current ethical review process in the U.S. towards a more collaborative and adaptive review process.

CHAPTER 8

Conclusion

Ethical review of research including human subjects is essential to ensuring the well-being of research participants and that research is conducted within according to regulations and current understanding of ethics in each particular research field. The latter in particular remains difficult, as ethical review today still largely follows a one-size-fits-all approach. Approaching this problem from a computer science perspective might not seem most obvious. However, as I hope to have convened in this work, a technological approach that takes into account the societal aspects, focusing on all human stakeholders involved allows to think differently about the current ethical review system and as such, can contribute by suggesting paths that have not been considered before, as the IRB experts interviewed in this work have also underlined.

Focusing on the example of ethical review for HC-based CS projects, this work has introduced, derived requirements of, developed the technical architecture and specification of, as well as empirically validated core details of CAER, a new approach to ethical review, with the aim of improving ethical review with respect to suitability, adaptivity, equity and efficiency. CAER goes beyond existing ethical review systems in allowing for dynamic and collaborative adaptation to the specific ethical considerations of different research fields, and improves the review process through semi-automatic evaluation of projects based on a rule engine and the adaptive ethical guidelines.

The HC-based approach allows for keeping humans in-the-loop during automated evaluation, while at the same time continuously refining the underlying rule set so that the platform keeps adapting and improving. The two core design aspects of collaborative guideline management and automated review were validated via user testing on proof-of-concept prototypes and IRB expert feedback.

The results of this evaluation suggest that, using CAER, the ethical review process can be made to be more efficient than established procedures and that its adaptive design makes CAER potentially more suitable for various domains and research fields than existing systems. Furthermore, the testing of the prototypes showed that no expert knowledge is required for using CAER, suggesting CAER may make the ethical review process more accessible while simultaneously increasing its transparency. This way, the new CAER system offers an appropriate way to consider the changing nature of ethics and values.

Beyond its aim of improving the ethical review process, CAER has the potential to play an important role in furthering a much-needed discussion on AI ethics and other emergent research fields and to function as an educational tool that makes ethical review more comprehensible and accessible to a broader public.

To fully explore the utility of CAER and to refine its new functionalities, further research and testing is required. Nevertheless, this presents a significant first step in laying the foundation for rethinking IRB processes and introducing a new collaborative and adaptive approach to ethical review.

APPENDIX A

Appendix

A.1 OpenAPI Schema definitions

The following tables include the full set of JSON schemas for the different objects returned and/or referenced by different API routes outlines in 5.2.3.

Path	#/components/schemas/DerivativeRule

Schema	{
	"title": "DerivativeRule",
	"required": [
	"title",
	"description",
	"rule_type",
	"rule_ids",
	"operator",
	"rule_id",
	"rule_status"
	],
	"type": "object",
	"properties": {
	"title": {
	"title": "Title",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"rule_type": {
	"title": "Rule Type",
	"enum": [
	"Derivative_rule"
	]
	},
	"rule_ids": {
	"title": "Rule Ids",
	"type": "array",
	"items": {
	"type": "integer"
	}
	},
	"operator": {
	"$ref": "#/components/schemas/OperatorType"
	},
	"rule_id": {
	"title": "Rule Id",
	"type": "integer"
	},
	"rule_status": {
	"$ref": "#/components/schemas/RuleStatus"
	}
	}
	}

Path	#/components/schemas/DerivativeRuleCreate

Schema	{
	"title": "DerivativeRuleCreate",
	"required": [
	"title",
	"description",
	"rule_type",
	"rule_ids",
	"operator"
	],
	"type": "object",
	"properties": {
	"title": {
	"title": "Title",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"rule_type": {
	"title": "Rule Type",
	"enum": [
	"Derivative_rule"
	]
	},
	"rule_ids": {
	"title": "Rule Ids",
	"type": "array",
	"items": {
	"type": "integer"
	}
	},
	"operator": {
	"$ref": "#/components/schemas/OperatorType"
	}
	}
	}

Path	#/components/schemas/DerivativeRuleUpdate

Schema	{
	"title": "DerivativeRuleUpdate",
	"type": "object",
	"properties": {
	"title": {
	"title": "Title",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"rule_ids": {
	"title": "Rule Ids",
	"type": "array",
	"items": {
	"type": "integer"
	}
	},
	"operator": {
	"$ref": "#/components/schemas/OperatorType"
	}
	}
	}

Path	#/components/schemas/EvaluationStatus

Schema	{
	"title": "EvaluationStatus",
	"enum": [
	"pending",
	"under_human_review",
	"approved",
	"rejected"
	],
	"description": "An enumeration."
	}

Path	#/components/schemas/OperatorType

Schema	{
	"title": "OperatorType",
	"enum": [
	"And",
	"Or",
	"Not"
	],
	"description": "An enumeration."
	}

Path	#/components/schemas/Project

Schema	{
	"title": "Project",
	"required": [
	"name",
	"description",
	"id",
	"user_id",
	"reviewer_user_id",
	"eval_status"
	],
	"type": "object",
	"properties": {
	"name": {
	"title": "Name",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"id": {
	"title": "Id",
	"type": "integer"
	},
	"user_id": {
	"title": "User Id",
	"type": "integer"
	},
	"reviewer_user_id": {
	"title": "Reviewer User Id",
	"type": "array",
	"items": {
	"type": "integer"
	}
	},
	"eval_status": {
	"$ref": "#/components/schemas/EvaluationStatus"
	}
	}
	}

Path	#/components/schemas/ProjectCreate

Schema	{
	"required": [
	"name",
	"description",
	"user_id"
	],
	"type": "object",
	"properties": {
	"name": {
	"title": "Name",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"user_id": {
	"title": "User Id",
	"type": "integer"
	}
	}
	}

Path	#/components/schemas/ProjectUpdate

Schema	{
	"type": "object",
	"properties": {
	"name": {
	"title": "Name",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	}
	}
	}

Path	#/components/schemas/Question

Schema	{
	"title": "Question",
	"required": [
	"question",
	"question_id",
	"question_status"
	],
	"type": "object",
	"properties": {
	"question": {
	"title": "Question",
	"type": "string"
	},
	"question_id": {
	"title": "Question Id",
	"type": "integer"
	},
	"question_status": {
	"$ref": "#/components/schemas/QuestionStatus"
	}
	}
	}

Path	#/components/schemas/QuestionCreate

Schema	{
	"required": [
	"question"
	],
	"type": "object",
	"properties": {
	"question": {
	"title": "Question",
	"type": "string"
	}
	}
	}

Path	#/components/schemas/QuestionRule

Schema	{
	"title": "QuestionRule",
	"required": [
	"title",
	"description",
	"rule_type",
	"question_id",
	"response_vals",
	"rule_id",
	"rule_status"
	],
	"type": "object",
	"properties": {
	"title": {
	"title": "Title",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"rule_type": {
	"title": "Rule Type",
	"enum": [
	"Question_rule"
	]
	},
	"question_id": {
	"title": "Question Id",
	"type": "integer"
	},
	"response_vals": {
	"$ref": "#/components/schemas/ResponseValue"
	},
	"rule_id": {
	"title": "Rule Id",
	"type": "integer"
	},
	"rule_status": {
	"$ref": "#/components/schemas/RuleStatus"
	}
	}
	}

Path	#/components/schemas/QuestionRuleCreate

Schema	{
	"title": "QuestionRuleCreate",
	"required": [
	"title",
	"description",
	"rule_type",
	"question_id",
	"response_vals"
	],
	"type": "object",
	"properties": {
	"title": {
	"title": "Title",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"rule_type": {
	"title": "Rule Type",
	"enum": [
	"Question_rule"
	]
	},
	"question_id": {
	"title": "Question Id",
	"type": "integer"
	},
	"response_vals": {
	"$ref": "#/components/schemas/ResponseValue"
	}
	}
	}

Path	#/components/schemas/QuestionRuleUpdate

Schema	{
	"title": "QuestionRuleUpdate",
	"type": "object",
	"properties": {
	"title": {
	"title": "Title",
	"type": "string"
	},
	"description": {
	"title": "Description",
	"type": "string"
	},
	"question_id": {
	"title": "Question Id",
	"type": "integer"
	},
	"response_vals": {
	"$ref": "#/components/schemas/ResponseValue"
	}
	}
	}

Path	#/components/schemas/QuestionStatus

Schema	{
	"title": "QuestionStatus",
	"enum": [
	"pending_for_review",
	"accepted",
	"rejected"
	],
	"description": "An enumeration."
	}

Path	#/components/schemas/QuestionUpdate

Schema	{
	"type": "object",
	"properties": {
	"question": {
	"title": "Question",
	"type": "string"
	}
	}
	}

Path	#/components/schemas/Response

Schema	{
	"required": [
	"project_id",
	"question_id",
	"response_value"
	],
	"type": "object",
	"properties": {
	"project_id": {
	"title": "Project Id",
	"type": "integer"
	},
	"question_id": {
	"title": "Question Id",
	"type": "integer"
	},
	"response_value": {
	"$ref": "#/components/schemas/ResponseValue"
	}
	}
	}

Path	#/components/schemas/ResponseValue

Schema	{
	"title": "ResponseValue",
	"enum": [
	"Yes",
	"No",
	"N/A"
	],
	"description": "An enumeration."
	}

Path	#/components/schemas/Rule

Schema	{
	"title": "Rule",
	"required": [
	"rule"
	],
	"type": "object",
	"properties": {
	"rule": {
	"title": "Rule",
	"anyOf": [
	{
	"$ref": "#/components/schemas/QuestionRule"
	},
	{
	"$ref": "#/components/schemas/DerivativeRule"
	}
	]
	}
	}
	}

Path	#/components/schemas/RuleCreate

Schema	{
	"required": [
	"rule"
	],
	"type": "object",
	"properties": {
	"rule": {
	"title": "Rule",
	"anyOf": [
	{
	"$ref": "#/components/schemas/QuestionRuleCreate"
	},
	{
	"$ref": "#/components/schemas/DerivativeRuleCreate
	"
	}
	]
	}
	}
	}

Path	#/components/schemas/RuleStatus

Schema	{
	"title": "RuleStatus",
	"enum": [
	"pending_for_review",
	"accepted",
	"rejected"
	],
	"description": "An enumeration."
	}

Path	#/components/schemas/RuleUpdate

Schema	{
	"required": [
	"rule"
	],
	"type": "object",
	"properties": {
	"rule": {
	"title": "Rule",
	"anyOf": [
	{
	"$ref": "#/components/schemas/QuestionRuleUpdate"
	},
	{
	"$ref": "#/components/schemas/DerivativeRuleUpdate
	"
	}
	]
	}
	}
	}

A.2 Additional example test cases for future work

A.2.1 Example case study 3: *Dream Catchers*

Dream Catchers is a HC-based CS project developed by the Human Computation Institute in partnership with the AI for Good Laboratory at Microsoft, the Cribs for Kids program and the Seattle Children's Hospital. Its purpose is to advance Sudden Infant Death Syndrome (SIDS) research which is especially important since "SIDS is the leading cause of death of infants 1 month to 1 year old in developed countries." [140] but it is not quite clear what causes SIDS, "the unexplained death of a seemingly healthy baby less than a year old" [140]. Despite this missing knowledge it is possible to reduce the risk of SIDS by practicing safe sleep positions. For example, a firm sleep surface should be used and infants should be placed on their back. With *Dream Catchers*, participants can help train a machine learning model to identify unsafe sleeping positions by labeling images of sleeping infants as "safe" or "unsafe" and by providing some further information on the specificity of the unsafe sleeping position (see figure number A.1. The project also informs participants about

safe sleep guidelines for infants. The long-term goal of the developers is to "help create powerful automatic tools to detect unsafe sleep, conduct parent behavior studies, identify best ways to train parents about safe sleep, and open various other research pathways" [140].

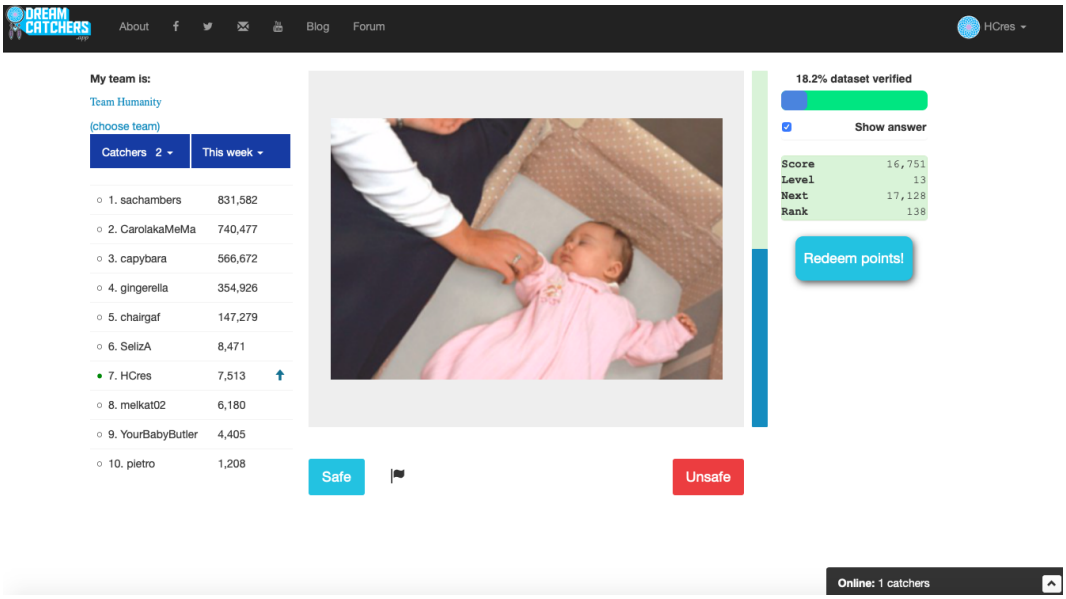


Figure A.1: *Dream Catchers* user interface before the introduction of the new features (Screenshot taken on November 21, 2019).

Dream Catchers is a derivative of *Stall Catchers*, the first CS game developed by the Human Computation Institute focusing on Alzheimer's disease research. After having been launched in October of 2019 and having run for about a year, *Dream Catchers* has been taken offline for a full relaunch on a new platform. While the first version was about labeling existing images of infants from free online sources to train machine learning models to distinguish safe from unsafe sleep conditions in images, the new to be launched version of *Dream Catchers* goes a step further. To develop these automatic tools, more training data annotated by human participants is needed. Therefore *Dream Catchers* is going to include a new feature that allows participants to upload their own images of their sleeping infants. Parents or legal representatives uploading images of their child have to make sure to not include any personally identifiable information in the images (such as name or date of birth of the infant). Additionally the infants' eyes have to be blurred by the uploading user. Before including the images in the platform they go through a vetting process in which trusted members of the community with special rights review if the images have been anonymized correctly. If three trusted curators have reviewed the image and consider it sufficiently anonymous the images will be released to the be analyzed by *Dream Catchers* participants.

To contribute participants must first agree to the Terms and Privacy policy of *Dream Catchers* to register and then must actively agree to a consent form to participate. They confirm that they have read the consent form by activating the checkbox "I agree to participate" and indicating their age. Participants have to be at least 18 years old and have to reside in the U.S. *Dream Catchers* collects participation data including images analyzed and time spent in the game. Participants allow the Human Computation Institute to make any use of these data but will always own their contribution. Personal information as well as

uploaded images will be stored anonymously in secure cloud-based storage.

To participate, a computer or mobile device with internet connection is necessary. *Dream Catchers* will be accessed via a web browser. Participation will not pose any risks to participants beyond any normal risks associated with playing an online game. There is no medical research being conducted on the participants. Parents or legal representatives may ask for their personal information and the uploaded images to be removed. However, they are informed in the Terms and Policy that third parties may have already downloaded the pictures at that point.

Once the machine learning model has been trained on sufficient *Dream Catchers* data, it will be publicly shared. Result will be shared with the participants via the Human Computation Institute's blog and the in-game chat. Participants will be acknowledged in publications related to Dream Catchers.

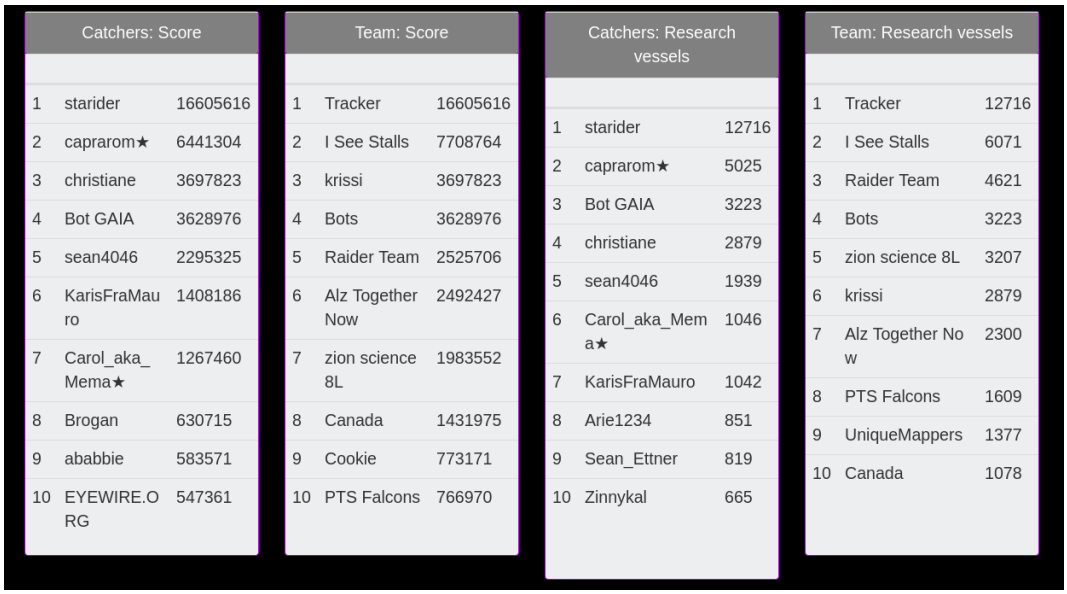
A.2.2 Example case study 4: AI-bots in *Stall Catchers*

Stall Catchers is an online CS project developed by the Human Computation Institute and launched in 2016 in which participants contribute to Alzheimer's disease research conducted at a biomedical engineering laboratory at Cornell University by analyzing research data presented in a game format. Participants are presented with short movie sequences depicting blood vessels in the brains of mice and are asked to decide whether a selected blood vessel is flowing or stalled. Combining humans and machines to solve this analysis problem has proven to be very effective for *Stall Catchers* [134]. This HC approach had been chosen because mere computational solutions such as with machine learning had previously proven not to meet the required data quality for research. In 2020 after having collected human annotation data in *Stall Catchers* for around four years, the Human Computation Institute had enough data to try a new computational approach again. In cooperation with the Human Computation Institute, DrivenData ran the machine learning competition, the "Clog Loss Challenge for Alzheimer's Research" [141]. They invited participants to build machine learning models to perform the task Stall Catchers participants perform, namely classifying blood vessels in 3D image stacks as flowing or stalled. The *Stall Catchers* crowd answers collected over the last years functioned as ground truth data for the models. More than 1,300 models were submitted to the challenge [141]. Although the created models do not outperform the human *Stall Catchers* crowd, they are very promising and allow to experiment with human-AI collaborations in *Stall Catchers* to facilitate the task humans have to perform and to make the analysis even more efficient and faster.

After the end of the challenge the winning models were made publicly available on Github [142] and the Human Computation Institute invited the winners and their models to work together with the institute. Laura Onac, the creator of the model winning the third place agreed to contribute and to help build an AI-bot based on her machine learning model. In the so-called bot-study the AI-bot GAIA will be introduced in *Stall Catchers* to annotate research data alongside humans [143].

The experiment is designed as follows: This experiment will last 24 hours in total and will be integrated in a "catchathon". A "catchathon" is a special *Stall Catchers* event in which participants working together in teams try to analyze as many research movies in *Stall Catchers* as possible and compete with other player groups. Extra leaderboards will be set up on which the GAIA bot's performance will also be displayed (see figure number A.2.

Participants will be informed about the bot study before the beginning of the catchathon via the Human Computation Institute's blog and the in-game chat. The GAIA bot will be recognizable on the leaderboard through the username "Bot GAIA". To participate participants must first agree to the Terms and Privacy policy of *Stall Catchers* to register but do



Catchers: Score

1	starider	16605616
2	caprarom★	6441304
3	christiane	3697823
4	Bot GAIA	3628976
5	sean4046	2295325
6	KarisFraMauro	1408186
7	Carol_aka_Mema★	1267460
8	Brogan	630715
9	ababbie	583571
10	EYEWIRE.ORG	547361

Team: Score

1	Tracker	16605616
2	I See Stalls	7708764
3	krissi	3697823
4	Bots	3628976
5	Raider Team	2525706
6	Alz Together Now	2492427
7	zion science 8L	1983552
8	Canada	1431975
9	Cookie	773171
10	PTS Falcons	766970

Catchers: Research vessels

1	starider	12716
2	caprarom★	5025
3	Bot GAIA	3223
4	christiane	2879
5	sean4046	1939
6	Carol_aka_Mema★	1046
7	KarisFraMauro	1042
8	Arie1234	851
9	Sean_Ettner	819
10	Zinnykal	665

Team: Research vessels

1	Tracker	12716
2	I See Stalls	6071
3	Raider Team	4621
4	Bots	3223
5	zion science 8L	3207
6	krissi	2879
7	Alz Together Now	2300
8	PTS Falcons	1609
9	UniqueMappers	1377
10	Canada	1078

Figure A.2: final Leaderboards of the catchathon 2021, @Human Computation Institute)

not have to give additional consent to participate in the bot study. By agreeing to the Terms and Privacy policy, participants consent to the collection and usage of their personal data and participation data. They can opt out at any time and request to have their personal information removed but their anonymized participation data will remain stored. After the bot study and the catchathon participants will be invited to share their experience with the bot on the Human Computation Institute forum. The results of the bot study will be published and shared with the participants via the Human Computation Institute blog and the in-game chat.

A.3 Ethical questions included in prototypes

1. Does your project collect information about individuals participating in your project (such as name, email address, age, phone number, etc.)?
2. Does the project collect the participant's geo location?
3. Will the geo location be encrypted/stored anonymously?
4. Does the project collect and store the Date of Birth of users?
5. Is your project a citizen science project?
6. Is the project an online citizen science project?
7. Does the project include pregnant women?
8. Do any communities around the research location have a negative attitude towards the conduct of research?
9. Will subjects or their representatives provide informed consent to take part in this research?
10. Will subjects or their representatives sign a written consent form?

11. Will subjects or their representatives actively accept the consent form (for example by clicking a checkbox)?
12. Does your project include children?
13. Is parental/legal representative consent required for the children to participate?
14. Will the results of the project be shared with the participants?
15. Will participants be acknowledged in publications for their contributions?
16. Does the project includes a machine learning model?
17. Is the trained model publicly shared?
18. Can participants opt out at any time?
19. Will personal information be deleted if participants decide to opt out?
20. Will you or others post the research on ClinicalTrials.gov?
21. Will you or others submit data from this research to the US Food and Drug Administration (FDA) or hold data from this research for inspection by the FDA?
22. Is the research a clinical trial of a device?
23. Does this research involve any form of human gene transfer as described in Section III-C of the NIH Guidelines?
24. Is the Principal Investigator (PI) a physician?
25. Does the Principal Investigator have a medical license?

A.4 Questionnaire - Validation of CAER

A.4.1 Questionnaire for IRB experts

- Would you feel comfortable to provide the following information: your age, the gender, you identify as, education, profession, IRB experience?
- What do you think about the current ethical review processes?
- In what ways does this current ethical review process and the ethical guidelines apply to emerging scientific fields such as citizen science and AI-research/HC? In what ways does it fall short?
- Our motivation to create a new platform for ethical review draws from the experiences we, the Human Computation Institute, made with applying for ethical review for HC-based citizen science projects. The problem we faced was that many of the questions asked in the review forms did not apply to our new online games for example, in which participants are not mere human subjects but also take over data analysis tasks etc. just like professional researchers.We realized that IRB was not fitted to citizen science and AI-research like HC and that there didn't exist common ethical guidelines for the ethical review of such projects because it is still an emergent field. In 2020 we ran a workshop on ethics in citizen science in which the need for another ethical review process became even more clear. My master thesis now forms the beginning of actually building a new ethical review platform that is more collaborative, efficient, equal, and adaptive to the specific fields of the projects to be reviewed. So, our motivation comes from experiences with IRB for new emergent fields and the goal is to improve ethical review for new emerging fields, we do not plan to change

ethical review altogether (though we hope that our approach will inspire others as well).

- How could the review process be improved to fit the needs, values and risks of such new emergent research fields including human subjects and/or research participants?
- To allow IRB experts to focus on those ethical questions to which there exists no answer in the current guidelines, we would like to outsource the initial review to an automated smart form that evaluates a project based on question-response pairs generated by IRB experts and trusted members of the scientific and participant community. This way, known ethical issues and questions could already be evaluated automatically before the human expert would have to take a look at the project to-be-reviewed. It would of course be guaranteed that a human expert would validate the automatically generated review solution at the end of the process and would address open questions. Where do you see the potential but maybe also the challenges with this approach?
- To make the process more efficient for project representatives who seek ethical review with their project, we plan to implement the automated evaluation system in a way that front-loads the most important questions, meaning those that most likely lead to a rejection of the project. The system would also only ask those questions that apply to the specific field of the research project. For example, it would not ask any details about the physical location of the study if the project were an online citizen science game. So, depending on the answers provided, the next question will be posed. This would not only speed up filling in the questionnaire for the project representative but would also make sure that all relevant questions and only those have been answered. Where do you see the potential but maybe also the challenges with this approach?
- In the literature on the current IRB process and the ethical guidelines that form the basis of ethical review today, it has been criticized by several scholars that these ethical guidelines are not up-to-date because they represent a fixed set of ethical issues but ethics and morals actually change over the years and have to be adapted to new technologies and societal changes. Moreover, it is criticized that these ethical guidelines have been tailored to the biomedical field which is why they do not respond to the needs, values, and risks of other scientific fields such as the humanities and, especially, new emergent fields such as AI-research and citizen science. With our new collaborative and adaptive ethical review system we want to invite IRB experts along with trusted members from impacted communities to update existing ethical guidelines and to suggest new questions or issues to focus on. These updates would then be presented to the other IRB experts and trusted members. Only if a broad enough representation of stakeholders agree with these updates, they will be accepted and included in the set of ethical issues and questions which also form the basis for the automated evaluation. With this functionality, the ethical guidelines would be expanded to support ethical review for specific new fields, in our case to HC-based citizen science, but could always be updated whenever experts agree on the need. How helpful would this functionality be for your work? Where do you see the potential but maybe also the challenges with this approach?
- What do you think about the role community representatives play today in IRB boards?
- Currently, community representatives are part of IRB boards, however, they do not necessarily have a lot to say due to power structures and hierarchies in the IRB boards etc. We think that especially for new emergent fields such as online citizen science and AI-research it is essential to include trusted representatives of the communities in the

development and "maintenance" of the ethical guidelines since the values and risks have not been sufficiently identified and discussed. By allowing trusted community representatives to suggest changes to existing ethical issues as well as to suggest new ones and to vote on suggestions by others, we want to make IRB review more fair and equal and broadly representative of values across stakeholders. What do you think about this approach? (advantages/potential , disadvantages/risks)

- Would you like to test the following prototypes yourself? a) smart evaluation form compared to classical form and b) interface to update rules and issues as well as to suggest new ones. Or would you prefer a walk-through presentation of the prototypes?
- What do you think about the prototypes? (Ask for both prototypes individually)
- Is there anything else we haven't discussed so far but which is very important to you?

A.4.2 Questionnaire for test users

Provided background information:
Thank you for participating in the testing process of two proof-of-concept prototypes for CAER, the new Collaborative and Adaptive Ethical Review Platform.

This research is about ethical review of citizen science and Artificial Intelligence (AI) research projects.

"Citizen Science" describes science projects that include the public to contribute to research by, for example, collecting or analyzing research data.

Ethical review describes the evaluation and oversight of research studies by ethical review boards. It ensures that research is conducted according to established ethical guidelines, regulations, and laws in order to protect the rights and well-being of human subjects as well as the integrity of the research process.

Motivation: The current ethical review system in the US has some challenges and does not always fit the needs of new emergent research fields such as citizen science and AI.

Aim: This research suggests a new ethical review platform that improves the process of ethical review for projects in these new emergent fields.

This introduction intentionally does not provide more information so as not to influence you before testing the proof-of-concepts.

In total, two proof-of-concept prototypes, each implementing one main feature of the new CAER system, should be tested. The process should not take more than 1 hour.

A.4.2.1 Questionnaire for prototype 1

- Would you feel comfortable to provide the following information: your age, the gender, you identify as, education, profession, IRB experience? Please read the example case study I sent you via email.
- Please go to the email I sent you and open the link for the first form [alternating form 1 or form 2]. Please fill in the form with the information provided on the example case study sheet.

- What do you think about the usability of the form you just filled in?
- How would you evaluate the procedure with respect to comprehensibility of the navigation on a scale from 1 to 5?
- What did you like about the process and interface?
- What could be improved?
- How would you evaluate the procedure with respect to efficiency of the overall experience on a scale from 1 to 10?
- Please estimate how many questions were relevant to evaluate the ethics of your project. (An estimate is sufficient)
- Do you have any other remarks you would like to share about the first form?
- For the second part of the first test, please fill in the following form based on the same example case study and information as the first one.
- What do you think about the usability of the form you just filled in?
- How would you evaluate the procedure with respect to comprehensibility of the navigation on a scale from 1 to 5?
- What did you like about the process and interface?
- What could be improved?
- How would you evaluate the procedure with respect to efficiency of the overall experience on a scale from 1 to 10?
- Please estimate how many questions were relevant to evaluate the ethics of your project. (An estimate is sufficient)
- Do you have any other remarks you would like to share about the second form and/or about the two forms in comparison to each other?

A.4.2.2 Questionnaire for prototype 2

The forms you just tested are based on so-called ethical issues that are created and maintained by ethical review experts and trusted community members. Ethical issues are questions such as "does your project collect personal information?" and answers may be "yes", "no", or "N/A". The system to be evaluated is meant to facilitate the process of adapting and keeping the ethical guidelines and issues up-to-date. This process is difficult in current approaches used by ethical review boards, which is why ethical guidelines that are currently applied to review research projects tend to be outdated and often do not fit the field of a specific research project.
The proposed system uses "rules" to determine whether a given project will be accepted or rejected, based on answers to the ethical questions. These rules may define, for example, that question x with answer "yes" combined with question y with answer "no" lead to a rejection. This rule would be expressed as
(x? $\rightarrow$ [yes] AND y? $\rightarrow$ [no] $\rightarrow$ reject).
It is possible to suggest modifications to these rules and to propose new rules as part of the overall system. I built a small proof-of-concept user interface to test the process of suggesting new ethical questions and rules. I would like to ask you to try to solve the following tasks using this new interface. After having completed all 6 tasks you will be asked to answer 9 questions based on your experience.

Completing the tasks and answering the questions should take around 20 minutes.

The following URL will bring you to the test user interface (Please note: You do not have to create your own account, see below):
https://airtable.com/appfSw8oMwTMs5XAb/pbd8BHf2hSwi0RmSl
Please use the following information to login:
email: [created for each test user]
password: [created for each test user]

Please try to solve the following tasks with the provided user interface, one after the other.

1. Suggest a new ethical question that asks if participants get acknowledged in project related publications. Answer options should include "yes", "no", and "N/A"
2. Update the ethical question "Will research subjects provide informed consent to take part in this research?" so that it includes not only subjects but also their representatives (representatives can also provide informed consent to take part in research).
3. Suggest a new rule where the question "Does the project collect the date of birth of users?" with the answer "yes" leads to the rejection of the project.
4. Suggest a new rule which combines the following existing rule with an existing question:
   rule: "(Does the project include a machine learning model? → [Yes])" AND question "Is the trained model publicly shared?" with answer "no" LEADS TO rejection
5. Update the following rule: "(Does the project include pregnant women? → [Yes]); leads to rejection" by disabling the existing rule and suggesting a new rule which states that the question "Is the project an online citizen science project?" with answer "no" and the question "Does the project include pregnant women?" with answer "yes" leads to the rejection of the project.
6. Update a rule: The following existing rule contains a mistake because it rejects projects that should be accepted: "(Is your project a citizen science project? → [Yes] AND Will the results of the project be shared with the participants? → [Yes])". Please correct this in a way so that a project will be rejected if it is a citizen science project and the results will not be shared with the participants: ("(Is your project a citizen science project? -¿ [Yes] AND Will the results of the project be shared with the participants? → [No])"

Finally, please answer 9 questions on your experience with the provided user interface:

- What do you think about the overall usability and understandability of the user interfaces for the tasks you performed?
- How would you evaluate the procedure with respect to comprehensibility of the navigation on a scale from 1 to 5?
- How would you evaluate the procedure in terms of clarity of how and where to perform each individual task?
- What did you like about the process and interface?
- What could be improved?
- How would you evaluate the procedure with respect to efficiency of the overall experience on a scale from 1 to 10?

- How comfortable would you feel to suggest new questions or rules in future with this system?
- How comfortable would you feel to update existing questions and rules in future?
- Do you see any potential modifications to ethical guidelines this system could not address?

Bibliography

[1] Deutsche Forschungsgemeinschaft. Guidelines for Safeguarding Good Research Practice. Code of Conduct. September 2019. URL: https://zenodo.org/record/3923602.

[2] Kyle B. Enfield and Jonathan Truwit. The Purpose, Composition, and Function of an Institutional Review Board: Balancing Priorities. *Respiratory Care*, 53(10):1330–1336, 2008.

[3] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. THE BELMONT REPORT. Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Technical report, April 1979. URL: https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf [Accessed: Sept 11, 2022].

[4] Kate Crawford. *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. Yale University Press, New Haven, 2021. OCLC: on1111967630.

[5] Mark A. Rothstein, John T. Wilbanks, and Kyle B. Brothers. Citizen Science on Your Smartphone: An ELSI Research Agenda: Currents in Contemporary Bioethics. *Journal of Law, Medicine & Ethics*, 43(4):897–903, 2015. URL: https://www.cambridge.org/core/product/identifier/S1073110500019227/type/journal_article.

[6] David B. Resnik. Citizen Scientists as Human Subjects: Ethical Issues. *Citizen Science: Theory and Practice*, 4(1):11, March 2019. URL: http://theoryandpractice.citizenscienceassociation.org/articles/10.5334/cstp.150/.

[7] Anne Bowser, Katie Shilton, Jenny Preece, and Elizabeth Warrick. Accounting for Privacy in Citizen Science: Ethical Research in a Context of Openness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 2124–2136, Portland, Oregon, USA, 2017. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2998181.2998305.

[8] David B. Resnik, Kevin C. Elliott, and Aubrey K. Miller. A framework for addressing ethical issues in citizen science. *Environmental Science & Policy*, 54:475–481, 2015. URL: https://www.sciencedirect.com/science/article/abs/pii/S1462901115001057?via%3Dihub.

[9] Kyle B. Brothers, Suzanne M. Rivera, R. Jean Cadigan, Richard R. Sharp, and Aaron J. Goldenberg. A Belmont Reboot: Building a Normative Foundation for Human Research in the 21st Century. *The Journal of Law, Medicine & Ethics*, 47(1):165–172, March 2019. URL: http://journals.sagepub.com/doi/10.1177/1073110519840497.

[10] Henry K. Beecher. Ethics and Clinical Research. *New England Journal of Medicine*, 274(24):1354–1360, June 1966. URL: http://www.nejm.org/doi/abs/10.1056/NEJM196606162742405.

[11] Won Oak Kim. Institutional review board (IRB) and ethical issues in clinical research. *Korean Journal of Anesthesiology*, 62(1):3, 2012. URL: http://ekja.org/journal/view.php?doi=10.4097/kjae.2012.62.1.3.

[12] Stacey A. Page and Jeffrey Nyeboer. Improving the process of research ethics review. *Research Integrity and Peer Review*, 2(1):14, December 2017. URL: http://researchintegrityjournal.biomedcentral.com/articles/10.1186/s41073-017-0038-7.

[13] Cornell University. Research with Human Participants. Institutional Review Board for Human Participant Research (IRB), 2019. URL: https://researchservices.cornell.edu/compliance/human-research [Accessed: Sept. 11, 2022].

[14] David B. Resnik. What Is Ethics in Research & Why Is It Important?, December 2020. URL: https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm [Accessed: Sept. 11, 2022].

[15] Libuse Hannah Veprek, Patricia Seymour, and Pietro Michelucci. Human Computation Requires and Enables a New Approach to Ethics. volume Proceedings of the Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation co-located with 34th Conference on Neural Information Processing Systems (NeurIPS 2020), pages 26–33, Vancouver, BC, Canada, 2020. URL: http://ceur-ws.org/Vol-2736/paper5.pdf [Accessed: Sept 11, 2022].

[16] Dietmar Hübner. *Einführung in die philosophische Ethik*. Number 4121 in utb Philosophie. Vandenhoeck & Ruprecht, Göttingen, 3., revised and corrected edition edition, 2021.

[17] Mordechai Haklay, Daniel Dörler, Florian Heigl, Marina Manzoni, Susanne Hecker, and Katrin Vohland. What Is Citizen Science? The Challenges of Definition. In Katrin Vohland, Anne Land-Zandstra, Luigi Ceccaroni, Rob Lemmens, Josep Perelló, Marisa Ponti, Roeland Samson, and Katherin Wagenknecht, editors, *The Science of Citizen Science*, pages 13–33. Springer International Publishing, Cham, 2021. URL: http://link.springer.com/10.1007/978-3-030-58278-4_2.

[18] Muki Haklay. Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In Daniel Sui, Sarah Elwood, and Michael Goodchild, editors, *Crowdsourcing Geographic Knowledge*, pages 105–122. Springer Netherlands, Dordrecht, 2013. URL: http://link.springer.com/10.1007/978-94-007-4587-2_7.

[19] Lea A. Shanley, Joey Hulbert, and Muki Haklay. CitSciDefinitions: Citizen Science Definitions, November 2019. URL: https://zenodo.org/record/3552753.

[20] Susanne Hecker, Mordechai Haklay, Anne Bowser, Zen Makuch, Johannes Vogel, and Aletta Bonn. *Citizen science: innovation in open science, society and policy*. 2018. URL: http://www.jstor.org/stable/10.2307/j.ctv550cf2 OCLC: 1132039153.

[21] Jonathan Silvertown. A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9):467–471, September 2009. URL: https://linkinghub.elsevier.com/retrieve/pii/S016953470900175X.

[22] Eric Betz. How a Christmas Tradition has Helped Track Billions of Vanishing Birds. *Discover magazine*, December 2020. URL: https://www.discovermagazine.com/planet-earth/how-a-christmas-tradition-has-helped-track-billions-of-vanishing-birds [Accessed: Sept. 11, 2022].

[23] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M. Harbison, Carl A. Fogarty, Joseph E. Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chemistry*, 13(7):651–659, July 2021. URL: http://www.nature.com/articles/s41557-021-00707-0.

[24] Anne Holohan. *Community, competition and citizen science: voluntary distributed computing in a globalized world*. Global connections. Ashgate, Farnham, Surrey ; Burlington, Vermont, 2013.

[25] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, June 2006. URL: https://www.wired.com/2006/06/crowds/ [Accessed: Sept. 11, 2022].

[26] Anthony Charles, Laura Loucks, Fikret Berkes, and Derek Armitage. Community science: A typology and its implications for governance of social-ecological systems. *Environmental Science & Policy*, 106:77–86, April 2020. URL: https://www.sciencedirect.com/science/article/pii/S1462901119300942.

[27] Darlene Cavalier and Eric B Kennedy. *Citizen science the rightful place of science*. Consortium for Science, Policy, & Outcomes, Arizona State University, 2016. OCLC: 1075065327.

[28] Lucy Danielle Robinson, Jade Lauren Cawthray, Sarah Elizabeth West, Aletta Bonn, and Janice Ansine. Ten principles of citizen science. In Susanne Hecker, Muki Haklay, Anne Bowser, Zen Makuch, J. Vogel, Johannes, and Aletta Bonn, editors, *Citizen Science: Innovation in Open Science, Society and Policy*, pages 27–40. UCL Press, London, 2017.

[29] François Bry, Clemens Schefels, and Christoph Wieser. Human computation. *it - Information Technology*, 60(1):1–2, March 2018. URL: http://www.degruyter.com/view/j/itit.2018.60.issue-1/itit-2018-0007/itit-2018-0007.xml.

[30] Edith Law and Luis Ahn. *Human computation*. Number 13 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2011.

[31] Luis Von Ahn. *Human Computation*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2005. UMI Order Number: AAI3205378.

[32] Alexander J. Quinn and Benjamin B. Bederson. Human Computation: A Survey and Taxonomy of a Growing Field. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–10, Vancouver, BC, Canada., May 2011. ACM.

[33] Pietro Michelucci. Human Computation and Convergence. In William Sims Bainbridge and Mihail C. Roco, editors, *Handbook of Science and Technology Convergence*, pages 455–474. Springer International Publishing, Cham, 2016. URL: https://doi.org/10.1007/978-3-319-07052-0_35.

[34] Martin Bogner, François Bry, Niels Heller, Stephan Leutenmayr, Sebastian Mader, Alexander Pohl, Clemens Schefels, Yingding Wang, and Christoph Wieser. Human Collaboration Reshaped: Applications and Perspectives. In Arndt Bode, Manfred Broy, Hans-Joachim Bungartz, and Florian Matthes, editors, *50 Jahre Universitäts-Informatik in München*, pages 47–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. URL: http://link.springer.com/10.1007/978-3-662-54712-0_4.

[35] Brian Koepnick, Jeff Flatten, Tamir Husain, Alex Ford, Daniel-Adriano Silva, Matthew J. Bick, Aaron Bauer, Gaohua Liu, Yojiro Ishida, Alexander Boykov, Roger D. Estep, Susan Kleinfelter, Toke Nørgård-Solano, Linda Wei, Foldit Players, Gaetano T. Montelione, Frank DiMaio, Zoran Popović, Firas Khatib, Seth Cooper, and David Baker. De novo protein design by citizen scientists. *Nature*, 570(7761):390–394, June 2019. URL: http://www.nature.com/articles/s41586-019-1274-4.

[36] zooniverse. Galaxy Zoo. n.d. URL: https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/about/results [Accessed: Sept 11, 2022].

[37] Pietro Michelucci. The People and Serendipity of the EyesOnALZ project. *When Citizens Do Science: Stories from Labs, Garages, and Beyond.*, 9(1):29–33, 2019.

[38] Oliver Bracko, Lindsay K. Vinarcsik, Jean C. Cruz Hernández, Nancy E. Ruiz-Uribe, Mohammad Haft-Javaherian, Kaja Falkenhain, Egle M. Ramanauskaite, Muhammad Ali, Aditi Mohapatra, Madisen A. Swallow, Brendah N. Njiru, Victorine Muse, Pietro E. Michelucci, Nozomi Nishimura, and Chris B. Schaffer. High fat diet worsens Alzheimer's disease-related behavioral abnormalities and neuropathology in APP/PS1 mice, but not by synergistically decreasing cerebral blood flow. *Scientific Reports*, 10(1):9884, June 2020. URL: https://www.nature.com/articles/s41598-020-65908-y number: 1 publisher: Nature Publishing Group.

[39] Chris Lintott and Jason Reed. Human Computation in Citizen Science. In Pietro Michelucci, editor, *Handbook of Human Computation*, pages 153–162. Springer New York, New York, NY, 2013. URL: http://link.springer.com/10.1007/978-1-4614-8806-4_14.

[40] Laura Jeanine Morris Stark. *Behind closed doors: IRBs and the making of ethical research*. Morality and society series. The University of Chicago Press, Chicago ; London, 2012.

[41] Manas Tungare, Ben Hanrahan, Ricardo Quintana-Castillo, Michael Stewart, and Manuel A Pérez-Quiñones. Collaborative Human Computation as a Means of Information Management. *Proceedings of the 2nd International Workshop on Collaborative Information Seeking at CSCW 2010*, 2010. URL: https://static.googleusercontent.com/media/research.google.com/de//pubs/archive/37450.pdf [Accessed: Sept 11, 2022].

[42] David T Lee. Modeling collaborative work in human computation. Vancouver, Canada, 2020. URL: https://tech4good.soe.ucsc.edu/assets/docs/neurips-coopai-2020-mo.pdf [Accessed: Sept. 11, 2022].

[43] Cayuse. Cayuse | eRA Software | Electronic Research Administration Systems. n.d. URL: https://cayuse.com/ [Accessed: Sept. 11, 2022].

[44] Axiom Mentor. Mentor Online IRB System. n.d. URL: https://19otva4t7392ycnen21gjwji-wpengine.netdna-ssl.com/wp-content/uploads/2020/08/Mentor-IRB-Detailed-Features-8-19.pdf [Accessed: Sept. 11, 2022].

[45] Tech Software. OneAegis (f/k/a IRBManager), 2022. URL: https://www.techsoftware.com/oneaegis [Accessed: Sept. 11, 2022].

[46] InfoReady. Software for Allocating Research Resources, 2022. URL: https://www.inforeadycorp.com [Accessed: Sept. 11, 2022].

[47] Key Solutions. IRB Software | Institutional Review Board | IRB Protocol, 2022. URL: https://www.keyusa.com/irb-software.html [Accessed: Sept. 11, 2022].

[48] Huron. Huron IRB Exchange - Huron, 2022. URL: https://www.huronconsultinggroup.com/insights/huron-irb-exchange [Accessed: Sept. 11, 2022].

[49] Maurice Collins. How One University Adopted InfoReady to Automate the IRB Approval Process, July 2018. URL: https://www.inforeadycorp.com/post/2018/07/25/how-one-university-adopted-infoready-review-to-automate-the-irb-approval-process [Accessed: Sept. 11, 2022].

[50] Advarra. Institutional Review Board Services | Advarra IRB Services, 2022. URL: https://www.advarra.com/review-services/institutional-review-board/ [Accessed: Sept. 11, 2022].

[51] Trial Interactive. Trial Interactive eClinical Solutions. n.d. URL: https://www.trialinteractive.com/ [Accessed: Sept. 11, 2022].

[52] Northwestern University. eIRB+: Institutional Review Board (IRB) Office - Northwestern University, 2022. URL: https://irb.northwestern.edu/submitting-to-the-irb/eirb/ [Accessed: Sept. 11, 2022].

[53] The University of Utah. ERICA, 2022. URL: https://irb.utah.edu/guidelines/erica-assistance/index.php [Accessed: Sept. 11, 2022].

[54] Nichelle Cobb, Elizabeth Witte, Maria Cervone, Aaron Kirby, Douglas MacFadden, Lee Nadler, and Barbara E. Bierer. The SMART IRB platform: A national resource for IRB review for multisite studies. *Journal of Clinical and Translational Science*, 3(4):129–139, August 2019. URL: https://www.cambridge.org/core/product/identifier/S2059866119003947/type/journal_article.

[55] Northwestern University. SMART IRB & IREx: Institutional Review Board (IRB) Office - Northwestern University. n.d. URL: https://irb.northwestern.edu/reliance/smart-irb-irex.html [Accessed: Sept. 11, 2022].

[56] Rachel L. Richesson, Jamie F. Malloy, Kathleen Paulus, David Cuthbertson, and Jeffrey P. Krischer. An Automated Standardized System for Managing Adverse Events in Clinical Research Networks. *Drug safety*, 31(10):807–822, 2008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6602073/ PMID: 18759506.

[57] Vanderbilt University Medical Center. IREx. n.d. URL: https://www.irbexchange.org/p/ [Accessed: Sept. 11, 2022].

[58] Stephen J. Rosenfeld. Managing What You Can't Measure—Institutional Review Board Decision Support Software. *Ochsner Journal*, 20(1):10–14, 2020. URL: http://www.ochsnerjournal.org/lookup/doi/10.31486/toj.19.0074.

[59] Sandra L. Schneider and Jane A. McCutcheon. Using an Automated Wizard to Process Minimal-Risk Research. *APS Observer*, 32, April 2019. URL: https://www.psychologicalscience.org/observer/using-an-automated-wizard-to-process-minimal-risk-research [Accessed: Sept. 11, 2022].

[60] Sandra L. Schneider and Jane A. McCutcheon. roof of concept: Use of a wizard for self-determination of IRB exempt status. Federal Demonstration Partnership, Washington D.C., 2018. URL: http://thefdp.org/default/assets/File/Documents/wizard_pilot_final_rpt.pdf [Accessed: Sept. 11, 2022].

[61] Jonathan Zong and J. Nathan Matias. Bartleby: Procedural and Substantive Ethics in the Design of Research Ethics Systems. *Social Media + Society*, 8(1):205630512210770, January 2022. URL: http://journals.sagepub.com/doi/10.1177/20563051221077021.

[62] Jonathan Zong. Designing and Evaluating Research Ethics Systems, February 2022. URL: https://citizensandtech.org/2022/02/designing-and-evaluating-research-ethics-systems/ [Accessed: Sept. 11, 2022].

[63] Karl Oder and Stephanie Pittman. The effect of computer automation on Institutional Review Board (IRB) office efficiency. *Research Management Review,*, 20(2):11, 2015. URL: https://files.eric.ed.gov/fulltext/EJ1063994.pdf [Accessed: Sept. 11, 2022].

[64] NeurIPS. 2022 Conference. URL: https://neurips.cc/ [Accessed: Sept. 11, 2022].

[65] Samy Bengio, Alina Beygelzimer, Kate Crawford, Jeanne Fromer, Iason Gabriel, Amanda Levendowski, Deborah Raji, and Marc'Aurelio Ranzato. Provisional Draft of the NeurIPS Code of Ethics. page 12, 2022. URL: https://openreview.net/pdf?id=zVoy8kAFKPr [Accessed: Sept. 11, 2022].

[66] Michael S. Bernstein, Margaret Levi, David Magnus, Betsy Rajala, Debra Satz, and Charla Waeiss. ESR: Ethics and Society Review of Artificial Intelligence Research. 2021. URL: https://arxiv.org/abs/2106.11521.

[67] North Star Review Board. North Star Review Board, June 2021. URL: https://learningirb.org/about/ [Accessed: Sept. 11, 2022].

[68] Tracey Brown, editor. *Peer review and the acceptance of new scientific ideas: discussion paper from a Working Party on equipping the public with an understanding of peer review; november 2002 - may 2004*. Sense About Science, London, 2004.

[69] Dale J. Benos, Edlira Bashari, Jose M. Chaves, Amit Gaggar, Niren Kapoor, Martin LaFrance, Robert Mans, David Mayhew, Sara McGowan, Abigail Polter, Yawar Qadri, Shanta Sarfare, Kevin Schultz, Ryan Splittgerber, Jason Stephenson, Cristy Tower, R. Grace Walton, and Alexander Zotov. The ups and downs of peer review. *Advances in Physiology Education*, 31(2):145–152, June 2007. URL: https://www.physiology.org/doi/10.1152/advan.00104.2006.

[70] Jeannette M. Wing and Ed H. Chi. Reviewing peer review. *Communications of the ACM*, 54(7):10–11, July 2011. URL: https://dl.acm.org/doi/10.1145/1965724.1965728.

[71] R. Smith. Peer review: reform or revolution? *BMJ*, 315(7111):759–760, September 1997. URL: https://www.bmj.com/lookup/doi/10.1136/bmj.315.7111.759.

[72] Nihar B. Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, June 2022. URL: https://dl.acm.org/doi/10.1145/3528086.

[73] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An Efficient Boosting Algorithm for Combining Preferences. *The Journal of Machine Learning Research*, 4:933–969, 2003.

[74] Ioannis Mitliagkas, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath. User rankings from comparisons: Learning permutations in high dimensions. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1143–1150, Monticello, IL, September 2011. IEEE. URL: http://ieeexplore.ieee.org/document/6120296/.

[75] Giorgio Maria Di Nunzio, Evangelos Kanoulas, and Prasenjit Majumder. Augmented Intelligence in Technology-Assisted Review Systems (ALTARS 2022): Evaluation Metrics and Protocols for eDiscovery and Systematic Review Systems. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, volume 13186 of *Lecture Notes in Computer Science*, pages 557–560. Springer International Publishing, Cham, 2022. URL: https://link.springer.com/10.1007/978-3-030-99739-7_69.

[76] EasyChair. EasyChair. n.d. URL: https://easychair.org/overview [Accessed: Sept. 11, 2022].

[77] Microsoft Corporation. Microsoft Conference Management Toolkit, 2022. URL: https://cmt3.research.microsoft.com/About [Accessed: Sept. 11, 2022].

[78] Simon Fraser University Library. Open Journal Systems | Public Knowledge Project, 2014. URL: https://pkp.sfu.ca/ojs/ [Accessed: Sept. 11, 2022].

[79] Scholastica. Scholastica: Academic journal publishing software and services. n.d. URL: https://scholasticahq.com/ [Accessed: Sept. 11, 2022].

[80] Hindawi Limited. Hindawi Home. n.d. URL: https://www.hindawi.com/ [Accessed: Sept. 11, 2022].

[81] OpenReview. OpenReview. n.d. URL: https://openreview.net/ [Accessed: Sept. 11, 2022].

[82] OpenConf. OpenConf Peer-Review, Conference and Abstract Management Software System. n.d. URL: https://www.OpenConf.com/ [Accessed: Sept 11, 2022].

[83] CERN. Indico - Home. n.d. URL: https://getindico.io/ [Accessed: Sept 11, 2022].

[84] Wikipedia. Main Page, June 2022. URL: https://en.wikipedia.org/w/index.php? title=Main_Page&oldid=1093586708 Page Version ID: 1093586708 [Accessed: Sept. 11, 2022].

[85] Wikipedia. Wikipedia: About, August 2022. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=1104085457 Page Version ID: 1104085457 [Accessed: Sept. 11, 2022].

[86] Desire Athow. Best online form builders of 2022, 2022. URL: https://www.techradar.com/best/best-online-form-builder [Accessed: Sept. 11, 2022].

[87] AirTable. Airtable, August 2022. URL: https://airtable.com/ [Accessed: Sept. 11, 2022].

[88] Airtable. How to create a form in Airtable. n.d. URL: https://support.airtable.com/hc/en-us/articles/360058735154-How-to-create-a-form-in-Airtable [Accessed: Sept. 11, 2022].

[89] Google. What can you do with Forms? - Google Workspace Learning Center, 2022. URL: https://support.google.com/a/users/answer/9302965?hl=en [Accessed: Sept. 11, 2022].

[90] Jotform Inc. Kostenlos Online Formulare Erstellen | Jotform. n.d. URL: https://www.jotform.com/ [Accessed: Sept. 11, 2022].

[91] Jotform. Smart Forms: Conditional Logic for Online Forms. n.d. URL: https://www.jotform.com/help/57-smart-forms-conditional-logic-for-online-forms/ [Accessed: Sept. 11, 2022].

[92] Typeform. Typeform: People-Friendly Forms and Surveys. n.d. URL: https://www.typeform.com/ [Accessed: Sept. 11, 2022].

[93] BERA (British Educational Research Association). Anticipating the application & unintended consequences of practitioner research. Technical Report 3, 2019. URL: https://www.bera.ac.uk/publication/anticipating-the-application-unintended-consequences-of-practitioner-research [Accessed: Sept. 11, 2022].

[94] Margaret R. Moon. The History and Role of Institutional Review Boards: A Useful Tension. *Virtual Mentor*, 11(4):311–316, 2009.

[95] Advisory Committee on Human Radiation Experiments. ACHRE Report. Technical report. n.d. URL: https://bioethicsarchive.georgetown.edu/achre/final/index.html [Accessed: Sept 11, 2022].

[96] Felix Khin-Maung-Gyi. The History and Role of Institutional Review Boards: Local and Central IRBs: A Single Mission. *Virtual Mentor*, 11(4):317–320, 2009.

[97] Office for Human Research Protections (OHRP). Federal Policy for the Protection of Human Subjects ('Common Rule'). Technical report, hhs.gov, March 2018. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html#:~:text=items%2C%20about%20Regulations-,Federal%20Policy%20for%20the%20Protection%20of%20Human%20Subjects%20('Common,of%20Biomedical%20and%20Behavioral%20Research. [Accessed: Sept 11, 2022].

[98] World Medical Association. WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects. Technical report, July 2018. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [Accessed: Sept 11, 2022].

[99] European Medicines Agency (EMA). Good clinical practice, September 2018. URL: https://www.ema.europa.eu/en/human-regulatory/research-development/compliance/good-clinical-practice [Accessed: Sept. 11, 2022].

[100] Office for Human Research Protections (OHRP) and U.S. Department of Health & Human Services. Excerpts from the January 19, 2017 Revised Common Rule Preamble. Technical report, August 2020. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/2018-req-preamble/index.html#46.108 [Accessed: Sept 11, 2022].

[101] Patricia Cohen. As Ethics Panels Expand Grip, No Field Is Off Limits. *The New York Times*, February 2007. URL: https://www.nytimes.com/2007/02/28/arts/28board.html [Accessed: Sept 11, 2022].

[102] James H. Sanders and Christine Ballengee-Morris. Troubling the IRB: Institutional Review Boards' Impact on Art Educators Conducting Social Science Research Involving Human Subjects. *Studies in Art Education*, 49(4), 2008. URL: http://www.jstor.com/stable/25475872 [Accessed: Sept 11, 2022].

[103] Louise-Anne McNutt. institutional review board. *Encyclopedia Britannica*, January 2019. URL: https://www.britannica.com/topic/institutional-review-board [Accessed: Sept. 11, 2022].

[104] Simon N. Whitney. *Balanced Ethics Review*. Springer International Publishing, Cham, 2016. URL: http://link.springer.com/10.1007/978-3-319-20705-6.

[105] Fatemeh Hajibabaee, Soodabeh Joolaee, Mohammad Ali Cheraghi, Pooneh Salari, and Patricia Rodney. Hospital/clinical ethics committees' notion: an overview. *Journal of Medical Ethics and History of Medicine*, 9(17):1–9, 2016.

[106] Ruth Macklin. How Independent Are IRBs? *IRB: Ethics & Human Research*, 30(3):15–19, 2008. URL: https://www.jstor.org/stable/30033260 [Accessed: Sept 11, 2022].

[107] Office for Human Research Protections (OHRP). Companion Q&As about the Revised Common Rule. Technical report, August 2018. URL: https://www.hhs.gov/ohrp/sites/default/files/Revised-Common-Rule-Q%26As-08-20-2018.pdf [Accessed: Sept 11, 2022].

[108] Office for Human Research Protections (OHRP). Lesson 4: Part 2: IRB Review. Technical report, June 2021. URL: https://www.hhs.gov/ohrp/education-and-outreach/online-education/human-research-protection-training/lesson-4-irb-review-of-research/index.html [Accessed: Sept 11, 2022].

[109] Office for Human Research Protections (OHRP). Human Subject Regulations Decision Charts: 2018 Requirements. Technical report, June 2020. URL: https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts-2018/index.html [Accessed: Sept 11, 2022].

[110] Cornell University. Does Your Project Require an Application to the Cornell IRB Office? Decision Tree #1. n.d. URL: https://researchservices.cornell.edu/sites/default/files/2019-05/IRB%20Decision%20Tree.pdf [Accessed: Sept. 11, 2022].

[111] Office for Human Research Protections (OHRP). Lesson 4: Independent Review of Research. Technical report, June 2021. URL: https://www.hhs.gov/ohrp/education-and-outreach/online-education/human-research-protection-training/lesson-4-irb-review-of-research/index.html [Accessed: Sept 11, 2022].

[112] n.a. The Complexity of the IRB Process: Some of the Things You Wanted to Know About IRBs but Were Afraid to Ask. *American Journal of Evaluation*, 26(3):353–361, September 2005. URL: http://journals.sagepub.com/doi/10.1177/109821400502600307.

[113] Office for Human Research Protections (OHRP). 45 CFR 46, March 2021. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html#46.110 [Accessed: Sept 11, 2022].

[114] Pietro Michelucci. The trouble with IRBs. Part I of an article series about Reinventing Ethical Review, August 2020. URL: https://medium.com/@pmichelu/the-trouble-with-irbs-f3f3dda82a87 [Accessed: Sept 11, 2022].

[115] C.K. Gunsalus, Edward M. Bruner, Nicholas C. Burbules, Leon Dash, Matthew Finkin, Joseph P. Goldberg, William T. Greenough, Gregory A. Miller, Michael G. Pratt, Masumi Iriye, and Deb Aronson. The Illinois White Paper: Improving the System for Protecting Human Subjects: Counteracting IRB "Mission Creep". *Qualitative Inquiry*, 13(5):617–649, July 2007. URL: http://journals.sagepub.com/doi/10.1177/1077800407300785.

[116] Steven Joffe. Revolution or Reform in Human Subjects Research Oversight. *Journal of Law, Medicine & Ethics*, 40(4):922–929, 2012. URL: https://www.cambridge.org/core/product/identifier/S1073110500017666/type/journal_article.

[117] Glenda Musoba, Stacy Jacob, and Leslie Robinson. The Institutional Review Board (IRB) and Faculty: Does the IRB Challenge Faculty Professionalism in the Social Sciences? *The Qualitative Report*, January 2015. URL: https://nsuworks.nova.edu/tqr/vol19/iss51/1/.

[118] Joseph A. Catania, Bernard Lo, Leslie E. Wolf, M. Margaret Dolcini, Lance M. Pollack, Judith C. Barker, Stacey Wertlieb, and Jeff Henne. Survey of U.S. Human Research Protection Organizations: Workload and Membership. *Journal of Empirical Research on Human Research Ethics*, 3(4):57–69, December 2008. URL: http://journals.sagepub.com/doi/10.1525/jer.2008.3.4.57.

[119] National Research Council. *Protecting participants and facilitating social and behavioral sciences research.* National Academies Press, Washington D.C., 2003.

[120] C. K. Gunsalus, Edward M. Bruner, Nicholas C. Burbules, Leon Dash, Matthew Finkin, Joseph P. Goldberg, William T. Greenough, Gregory A. Miller, and Michael G. Pratt. Mission Creep in the IRB World. *Science*, 312(5779):1441–1441, June 2006. URL: https://www.science.org/doi/10.1126/science.1121479.

[121] Zachary M. Schrag. ETHICAL TRAINING FOR ORAL HISTORIANS. *Perspectives on History*, 45(3), March 2007. URL: https://www.historians.org/publications-and-directories/perspectives-on-history/march-2007/ethical-training-for-oral-historians [Accessed: Sept 11, 2022].

[122] Yvonna S. Lincoln and William G. Tierney. Qualitative Research and Institutional Review Boards. *Qualitative Inquiry*, 10(2):219–234, April 2004. URL: http://journals.sagepub.com/doi/10.1177/1077800403262361.

[123] Tara Star Johnson. Qualitative Research in Question: A Narrative of Disciplinary Power With/in the IRB. *Qualitative Inquiry*, 14(2):212–232, March 2008. URL: http://journals.sagepub.com/doi/10.1177/1077800407308821.

[124] Monica Leisey. Qualitative Inquiry and the IRB: Protection at all Costs? *Qualitative Social Work*, 7(4):415–426, December 2008. URL: http://journals.sagepub.com/doi/10.1177/1473325008097138.

[125] Christopher Santos-Lang. Making it Easier to Make Your Own IRB. *Narrative Inquiry in Bioethics*, 9(1):E9–E12, 2019. URL: https://muse.jhu.edu/article/722817.

[126] Pietro Michelucci and Libuse Hannah Veprek. Toward Reinventing IRB for Citizen Science - a community discussion. Side event at the ECSA 2020 conference. September 2020. URL: https://www.youtube.com/watch?v=1V8y5y3XrRc [Accessed: Sept 11, 2022].

[127] Bernice S Elger and Eva De Clercq. Returning Results: Let's Be Honest! *Genet Test Mol Biomarkers.*, 21(3):134–139, 2017.

[128] Lisa M. Rasmussen, Anne Bowser, Caren Cooper, and Kathleen Lowenstein. Filling the 'Ethics Gap' in Citizen Science. A workshop report. Technical report, 2017. URL: https://www.niehs.nih.gov/research/supported/translational/peph/webinars/ethics/rasmussen_508.pdf [Accessed: Sept. 11, 2022].

[129] Mike Cohn. *User Stories: für die agile Software-Entwicklung mit Scrum, XP u.a.* mitp, Heidelberg München Landsberg Frechen Hamburg, 1. aufl edition, 2010.

[130] Bill Wake. INVEST in Good Stories, and SMART Tasks, 2003. URL: https://xp123.com/articles/invest-in-good-stories-and-smart-tasks/ [Accessed: Sept. 11, 2022].

[131] Glenn E. Krasner and Stephen T. Pope. A Description of the Model-View-Controller User Interface Paradigm in the Smalltalk-80 System, 1988. URL: http://heaveneverywhere.com/stp/PostScript/mvc.pdf [Accessed: Sept 11, 2022].

[132] Avraham Leff and James T Rayfield. Web-Application Development Using the Model/View/Controller Design Pattern. IBM Research Report RC 22002 (98835), March 2001. URL: https://dominoweb.draco.res.ibm.com/reports/rc22002.pdf [Accessed: Sept. 11, 2022].

[133] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making, August 2022. URL: http://arxiv.org/abs/2208.07960 arXiv:2208.07960 [cs].

[134] Oliver Bracko, Lindsay K. Vinarcsik, Jean C. Cruz Hernández, Nancy E. Ruiz-Uribe, Mohammad Haft-Javaherian, Kaja Falkenhain, Egle M. Ramanauskaite, Muhammad Ali, Aditi Mohapatra, Madisen Swallow, Brendah N. Njiru, Victorine Muse, Stall Catchers contributors, Pietro E. Michelucci, Nozomi Nishimura, and Chris B. Schaffer. High fat diet worsens pathology and impairment in an Alzheimer's mouse model, but not by synergistically decreasing cerebral blood flow. preprint, Neuroscience, December 2019. URL: http://biorxiv.org/lookup/doi/10.1101/2019.12.16.878397.

[135] Human Computation Institute. CrowdMeter. n.d. URL: https://crowdmeter.app/ [Accessed: Sept. 11, 2022].

[136] Pietro Michelucci, Yrjö T Gröhn, Alison Lynn Hill, John Krumm, John S L Parker, and Markus Schläpfer. CrowdMeter Results. n.d. URL: https://crowdmeter.app/wp-content/uploads/2020/10/cm_two_pager_r2.pdf [Accessed: Sept. 11, 2022].

[137] Human Computation Institute. Know Where the Crowds Are With the New CrowdMeter App, March 2022. URL: https://www.globenewswire.com/en/news-release/2022/03/23/2408931/0/en/Know-Where-the-Crowds-Are-With-the-New-CrowdMeter-App.html [Accessed: Sept. 11, 2022].

[138] Katja Nelson. Getting Started: Airtable Interface Designer | Airtable Guides. n.d. URL: https://www.airtable.com/guides/collaborate/getting-started-with-interface-designer [Accessed: Sept. 11, 2022].

[139] WCG IRB. Initial Review Submission Form. HRP-212. v 4.0.2.23, February 2020.

[140] Egle (Seplute). Stall Catchers had a baby! Introducing: Dream Catchers, October 2019. URL: https://blog.hcinst.org/introducing-dream-catchers/ [Accessed: Sept 11, 2022].

[141] Greg Lipstein. Meet the winners of the Clog Loss Challenge for Alzheimer's Research, September 2020. URL: https://www.drivendata.co/blog/clog-loss-alzheimers-winners [Accessed: Sept. 11, 2022].

[142] drivendata.org. Clog Loss: Advance Alzheimer's Research with Stall Catchers, September 2021. URL: https://github.com/drivendataorg/clog-loss-alzheimers-research [Accessed: Sept. 11, 2022].

[143] Grete Vaicaityte. Bots, that are going to play Stall Catchers along humans., September 2021. URL: https://blog.hcinst.org/bots-that-are-going-to-play-stall-catchers-along-humans/ [Accessed: Sept. 11, 2022].