Designing an Artificial Intelligence Research Review Committee

Sara R. Jordan
Chair, Center for Public Administration & Policy
Virginia Tech
900 N. Glebe Road, Arlington, VA 22203

email: srjordan@vt.edu
phone: +1-571-858-3061
skype: sara.r.jordan

PUBLISHED BY

KEYWORDS

- Ethics
- Research Ethics
- Artificial Intelligence
- Autonomous Systems
- Institutional Biosafety Committee

ABSTRACT

Calls for a review committee dedicated to ethical oversight of artificial intelligence research have not yet included serious considerations of the design of this committee. Here, a proposal for design of an artificial intelligence review committee review board is developed drawing upon the history and structure of existing research review committees, such as Institutional Review Boards (IRB), Institutional Animal Care and Use Committees (IACUC), and Institutional Biosafety Committees (IBCs). Drawing upon the risk-adjusted levels of review, a structure for evaluating the risk of AI projects is proposed. The review board structure recommended follows that of the IBC but with a blend of features from human subjects and animal care and use committees in order to improve implementation of risk-adjusted oversight mechanisms.

Designing an Artificial Intelligence Research Review Committee

INTRODUCTION

Calls for establishment of review boards for artificial intelligence (including: artificial general intelligence, artificial superintelligence, machine learning, and autonomous systems) ask these potential boards to review the ethics of AI research and technological development in the context of the uncertainties, risks, and benefits generated from these new technologies (Bostrom 2003, 2012; Eden, Moor, Soraker and Steinhard, 2013; Statt 2019, Stone 2018, Todd, 2019, Yampolskiy and Fox, 2013; Yudkowsky, 2008). These many calls bring the issue into stark relief but fail to address the pragmatic task of designing a review board (although see Sandler, Basl and Tiell 2019 for some preliminary work). The starting point to this article is that resolutions to problems raised by various authors requires that these conversations move beyond discussion of needs and ideas to address the design of such a committee. This article explicitly addresses the task of describing and then recommending design options for committees tasked to review artificial intelligence (AI) research.

Calo (2010, 2014, 2015), Marchant and Wallach (2015), and Calvo and Peters (2018) each make an academic case for ethical oversight of robotics and artificial intelligence research. Calo suggests that, in the absence of the now-defunct Office of Technology Assessment, a Federal Robotics Commission should be established to govern research involving "software that can harm" (Calo 2014). Marchant and Wallach suggest that a Governance Coordination Committee would look beyond solely "scientific and risk management aspects of technology" to "give greater emphasis to the variety of governmental and nongovernmental oversight and governance mechanisms that are in place or have been proposed" for governance of "nanotechnology, biotechnology, synthetic biology, applied neuroscience, geoengineering, regenerative medicine, robotics, and artificial intelligence" (2015). Concerns about "technology run amok", of "grey goo", and of Artificial Intelligence researchers being out of step with public concerns due to their professional self-interest echo those concerns that motivated creation of existing review institutions (Giles 2004; Hartzog 2015). Each of these pieces suggest that oversight is needed, which is a sentiment echoed in other venues, such as the recent Global Initiative on Ethics of Autonomous and Intelligent Systems (Amsterdam 2016; The IEEE Global Initiative 2019).

What these myriad calls for review do not do is to propose a review board structure. The specific design questions that must be asked include: What form and level of oversight should govern AI research? Should one board review all types of AI research, from natural language programming to artificial superintelligence? Should such a review board be organized at the local, national, international, or supranational level? Which professions and individuals should be counted among the actors responsible for reviewing ethics in these advanced technological fields? Should these boards have the power to stop research as other review boards do? What ought to be the domain of responsibility for such a committee when research interacts with other, existing research boards? With whom should the "buck stop" if research goes awry?

In the development of this article, I will first discuss the public concerns that motivates the call for an AIRC and relate these points to similar calls for formation of the common existing committees, the IRB, ULAC, and IBC.[1] Next, I will discuss the types of oversight proposed for human subjects, animal, and rDNA research and then discuss how these organizations form structures of oversight that could work to alleviate the motivating concerns for the Artificial Intelligence Review Committee or AIRC. Within this, I will discuss how traits of each review system could be incorporated into an AIRC system. Finally, I will offer some suggestions regarding which levels of political organization—global, national, local—these boards and officials should be arranged.

A BRIEF DESCRIPTION OF RESEARCH REVIEW

Researchers engaged in medical, social and behavioral, and animal research are familiar with the professional obligation to submit protocols for ethical review of their research before initiating work. In conversations about review boards for Artificial Intelligence, "IRBs" are often proposed as an institution to perform this task. Although IRBs are a recognizable institutional form, and as human subjects' research review is frequently discussed in training about research ethics, the background of the establishment of human subjects' review boards is not understood widely. In this section, review boards will be described to familiarize unfamiliar audiences with the role and powers of review boards and to outline the unique powers of these boards (Bankert and Amdur 2006; Edgar and Rothman 1995; Newgard 2002).

Human Subjects Research

The quick march of history obscures the fact that review boards are a recent addition to the research landscape. Human subjects review boards have been in operation for almost 65 years, with impressive spread to almost all institutions only within the past 35 years. Prior to the establishment of review boards, medical research proceeded through the non-systematic investigations of individual physicians working to address specific treatment concerns.

The earliest formal committees appeared in the NIH before World War II and were specifically designed to review high risk studies in fields such as cancer research (Vollman and Winau 1996). Three events motivated public calls for a robust public system of review boards. First, the revelations of crimes against humanity conducted in the name of "research" by Nazi and Japanese officials spurred development of the Nuremburg Code (first edition 1946), which outlined the foundational terms for ethical use of human subjects in research. Second, the Declaration of Helsinki (first edition 1948, latest edition 2013) set out further ethical terms and solidified the need for research review boards (World Medical Association 2013). Third, Henry Beecher's article that detailed multiple studies of dubious value and high degrees of harm brought examples of research needing review into the professional discussion (Beecher 1966, 1976).

The Nuremburg Code started with the (then) profound statement that: "The voluntary consent of the human subject is absolutely essential", which set the stage for one of the major actions of contemporary

1 As of September 1, 2019, there are 11,913 IRBs; 1,152 US domestic IACUCs and 339 non-US institutions with approved assurances for an IACUC; and 1362 IBCs registered with the Office for Human Research Protections, Office of Laboratory Animal Welfare, and NIH Office of Science Policy, respectively.

human subjects review boards—review of informed consent documents. The Nuremburg Code also set out the implicit need for scientific credential review and, in point 10, stipulated the expectation that ethical conduct requires a "scientist in charge" who has a more neutral perspective on the research and will put a stop to any work that is "likely to result in injury, disability, or death to the experimental subject" (Office of Human Research Protections, "Nuremburg Code", 2016).

Following on the principled statements of the Nuremburg Code, the Declaration of Helsinki outlined how review boards translate principles to practice in order to advance ethical research. The first Declaration (1964) described review boards in limited terms that expanded into descriptions in the latest edition (2013). Article 23 of the 2013 Declaration stipulates the basic obligations of review boards and researchers working with human subjects:

> "The research protocol must be submitted for consideration, comment, guidance and approval to the concerned research ethics committee before the study begins. This committee must be transparent in its functioning, must be independent of the researcher, the sponsor and any other undue influence and must be duly qualified. It must take into consideration the laws and regulations of the country or countries in which the research is to be performed as well as applicable international norms and standards but these must not be allowed to reduce or eliminate any of the protections for research subjects set forth in this Declaration. The committee must have the right to monitor ongoing studies. The researcher must provide monitoring information to the committee, especially information about any serious adverse events. No amendment to the protocol may be made without consideration and approval by the committee. After the end of the study, the researchers must submit a final report to the committee containing a summary of the study's findings and conclusions."

Which research should be reviewed, however, was brought into relief through the research of Beecher, who outlined numerous medical experiments in various clinical disciplines that put subjects at undue risk of harm, most frequently without disclosing the risks or without attending to study findings-in-progress that demonstrated research-related injuries to patients. Importantly Beecher's work highlighted that the substantive public interest in medical research can only be met by "intelligent, informed, conscientious, compassionate, responsible investigators" (1966, 33).

These three events set the stage for human subjects review, but four additional developments in the United States institutionalized review of human subjects of research: 1) revelation of the Tuskegee syphilis experiments, 2) development of the Belmont Report, 3) codification of the Common Rule (aka 45CFR46), and 4) establishment of the Federal Wide Assurance (FWA) system. Respectively, these four developments institutionalized responses to specific public concerns: 1) use of vulnerable individuals in research, 2) lack of systematic principles for governance of human subjects research, 3) lack of legal and regulatory frameworks to compel instantiation of ethical norms, and 4) unevenness in the application of principles and rules for research conducted with public funds.

The Tuskegee experiments highlighted a public fear that medical researchers could operate outside of law or conventional morality to satisfy questions of curiosity with invaluable human lives. Revelation of the Tuskegee syphilis experiments documented a 40 year "study" of the progress of a disease with well-established clinical markers in a racially-marginalized, functionally

and scientifically illiterate population, with no outside evaluation of the study, no establishment of subject safeguards, and no compensation for the participants. The Tuskegee revelation made explicit the need to consider vulnerable populations' needs within the ambit of research review and the need for a clear, scientifically sound and ethically principled approach to research conduct (Brandt 1978; Thomas and Quinn 1991).

Following on the heels of the Tuskegee revelation, from 1976 to 1978, a group of ethicists, physicians, laboratory researchers, lawyers and others gathered together in the Belmont, Maryland conference center to debate core principles to govern human subjects research. While the full proceedings of their debates appear in a two-volume examination of principles and alternatives, the public-facing Belmont Report is a brief document that makes clear that the following three principles form the bedrock of ethical review of human subjects research: justice, beneficence, and respect for persons (National Commission 1978; Office for Human Research Protections, "The Belmont Report", 2016).

The Belmont Report outlined recommendations for review according to ethical principles, *not* legally enforceable rules for review. The establishment of enforceable rules for review—including assignment of the Belmont Report as the modal document for research ethics—would come with passage of the "Common Rule" (45CFR46). The Common Rule is a cross-agency, regulatory document outlining the procedural, organizational, and ethical requirements for review of human subjects research, including research with vulnerable individuals, such as prisoners (Office of Human Research Protections, "Federal Policy for the Protection of Human Subjects", 2016). Originally passed in 1991, the Common Rule codified the steps of research review for projects sponsored by 18 different US federal agencies and departments, whether the research is conducted in the US or abroad. This document laid out critical definitions, organizational structures, and procedural review points as detailed in Box 1 below.

Significantly, the Common Rule defined research in a narrow sense that does not cover all human interaction. Research is a systematic investigation that is designed to contribute to generalizable knowledge about human physiology, pathology, psychology, or sociology, and so forth. The Common Rule codified the requirements for research review under the Declaration of Helsinki, requiring researchers to prepare a protocol document which details the purpose of the research, reason for sampling subject groups, methods of intervention or interaction, processes and forms that will be used to obtain consent from the participants, plans for rendering data confidential or anonymous, and plans for sharing, retaining and disseminating the research data. Once a protocol is submitted, it is reviewed to determine if the research poses a low or high level of risk. Under the terms of the Common Rule, IRBs may determine that the research is sufficiently innocuous that it is exempt from full review or is sufficiently concerning as to require full board review. Regardless of the level of review scrutiny, the requirement to obtain IRB review is imperative: failure to gain IRB approval may result in a researcher being charged with research misconduct or their publications being denied for review or retracted (Simes 1986; COPE 2011).

The Common Rule requires that protocols be reviewed by individuals with expert knowledge of the subject and knowledge of the local context of the research subjects involved. To assure the public that research in one location was not "safer" than research in another, the Office of Human Research Protections developed the Federal Wide Assurance system (Grady 2009, Newgard 2002). As described by the OHRP:

> Under an FWA, an institution commits to HHS that it will comply with the requirements set forth in 45 CFR part 46, as well as the Terms of Assurance. FWAs also are approved by OHRP for federalwide use, which means that other federal departments and agencies that have adopted the Federal Policy for the Protection of Human Subjects (also known as the Common Rule) may rely on the FWA for the research that they conduct or support... There is a single version of the FWA and the Terms of Assurance for domestic (U.S.) institutions and international (non-U.S.) institutions" (OHRP "Register IRBs and FWAs", 2016).

Human subjects review is not confined to the knowledge of local boards alone; institutions are supported by a host of non-governmental actors that include non-profit and advocacy organizations, such as Public Responsibility in Medicine & Research (PRIM&R), accreditation bodies such as the Association for the Accreditation of Human Research Protection Programs (AAHRPP), and education and certification bodies like the Association of Clinical Research Professionals (ACRP). Each of these institutions support the transmission of information, services, and best practices between actors invested in the widespread layers of human subjects review (Frumkin and Galaskiewicz 2004).

Animal Care and Use

There is a long history of animal use in research, particularly medical research for human benefit (Rollin 2006). In fact, the Declaration of Helsinki (article 21) makes it clear that:

> "Medical research involving human subjects must conform to generally accepted scientific principles, be based on a thorough knowledge of the scientific literature, other relevant sources of information, and adequate laboratory and, as appropriate, animal experimentation. The welfare of animals used for research must be respected."

Public rejection of animal research (e.g., vivisection) is long standing as is the demand for careful, expert, review of animal research (Cottingham 1978). Early cases of public outrage include the "Brown Dog" affair in the UK, revulsion at the University of California at Riverside blindness perception studies, and public outcry following a 1966 *Life Magazine* article comparing animal research facilities to concentration camps. Public concerns about review and oversight of animal use in research can be broken into three topics: 1) restriction on the use of companion animals in research; 2) the need for veterinarian informed oversight; and 3) the need to maintain a high quality of life and pain-free death for research animals (Carbone 2004).

Between 1960 and 1963, concerned veterinarians in Chicago organized animal care and use review boards, authored the well-known *Guide for the Care and Use of Laboratory Animals*, and organized a high level accreditation board that would become the present day Association for the Assessment and Accreditation of Laboratory Animal Care International (AAALAC). Around the same time, Congressional representatives drafted and passed the 1966 "Animal Welfare Act". The Animal Welfare Act, which has been amended nine times since, catalyzed relevant agencies to create tighter regulatory systems governing sourcing of animals, care of animals in research, and organization of Institutional Animal Care and Use Committee (IACUC) boards. Under this legal authority, the Public Health Service drafted policies on "Humane Care and Use of Laboratory Animals", the USDA drafted "Standard F" addressing the problem of animals stolen by unscrupulous Class B dealers, and the Public Health Service created an assurance program that requires institutions conducting animal research to file an Animal Welfare Assurance of Compliance with the Office of Laboratory Animal Welfare (OLAW) (Demers et al 2006; Interagency Research Animal Committee 2010; USDA "Animal Welfare Act" no date).

BOX 1: Policies and Procedures for Institutional Review Boards as Designated in the Common Rule (45CFR46, subpart A)

- IRBs must be registered with the Office of Human Research Protections, located in the Department of Health and Human Services (DHHS)

- An "institutional official" in each organization has power to obligate the institution to Common Rule requirements. IRBs registered with OHRP must have:
  - "At least five members, with varying backgrounds to promote complete and adequate review of research activities commonly conducted by the institution... If an IRB regularly reviews research that involves a vulnerable category of subjects... consideration shall be given to the inclusion of individuals who are knowledgeable and experienced in working with these subjects."
  - The members of an IRB should be broadly representative, non-discriminatory, include one "community representative" that is unaffiliated with the organization, and exclude individuals with conflicts of interest that would prevent fair review of the protocols.

- The members of the IRB should ensure that, except when waived by the IRB, "information [be] given to subjects as part of informed consent". Informed consent requires 8 pieces of information: "1. a statement that the study involves research, 2. an explanation of the purposes of the research and the expected duration of the subject's participation, 3. a description of the procedures to be followed, and identification of any procedures which are experimental", 4. "A description of any reasonably foreseeable risks....", 5. "a description of any benefits to the subject or others which may be reasonably expected", 6. disclosure of any alternative treatments or therapies, a statement detailing confidentiality of records, 7. "An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights, and whom to contact in the event of a research- related injury to the subject; and 8. A statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled, and the subject may discontinue participation at any time without penalty or loss of benefits to which the subject is otherwise entitled".

- Criteria for IRB approval of research includes:
  - "Risks to subjects are minimized...", "risks to subjects are reasonable in relation to anticipated benefits...", "selection of subjects is equitable...", and "informed consent will be sought [and] appropriately documented".

While the regulatory environment for care and use of animals focuses largely standards for research animal holding, housing, transfer, and euthanasia, the ethical environment for animal use in research addresses animal welfare from the start of research projects through their conclusion. The well-known 3R's of animal care and use—reduce, replace, and refine—reflect this start to finish perspective:

> "The guiding principles underpinning the humane use of animals in scientific research are called the three Rs. Any researcher planning to use animals in their research must first show why there is no alternative and what will be done to minimize numbers and suffering, i.e.: Replace the use of animals with alternative techniques, or avoid the use of animals altogether; Reduce the number of animals used to a minimum, to obtain information from fewer animals or more information from the same number of animals. Refine the way experiments are carried out, to make sure animals suffer as little as possible. This includes better housing and improvements to procedures which minimize pain and suffering and/or improve animal welfare" (Understanding Animal Research UK 2016).

Changes in knowledge about animal behavior, increasing public pressure, particularly from coordinated groups such as People for the Ethical Treatment of Animals (PETA) and the Humane Society, and advancements in veterinary care of laboratory animals pushed comparatively rapid policy changes in the governance of animal use in research, such as greater attention to appropriate habitat and interaction (see Kilkenny et al, 2010 and Mendelson 1996). Also, the incorporation of veterinarians into active roles in Institutional Animal Care and Use (IACUC) committees was a key change in the institutionalization of research animal welfare. Veterinarians included on committees filled powerful roles to ensure respectful use, quality of life, inclusion of socially and intellectually stimulating enclosures and increased opportunities for exercise and play for companion animals. Preservation of a dignified and reasonably pain free life extended to refined and specific rules regarding the use of analgesia, anesthesia, survival surgery, and restrictions on the type and duration of any introduction of pain and noxious stimuli (Silverman, Sockow and Murthy 2014; Wolfensohn and Lloyd 2008). Some examples of these rules are highlighted in Box 2, "Abbreviated US Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training".

The history of animal care and use committees have a stronger history of accreditation and supervision of animal use that is seen in human subjects use. For example, the human subjects review board accreditation program (AAHRPP) is less powerful and has a less global reach than the animal care and use accreditation body (AAALAC) (Resnik 2009, Rodriguez et al 2015). The AAALAC (Assessment and Accreditation of Laboratory Animal Care, International) draws standards of performance and care in animal research from the US, UK and EU, including the *Guide for the Care and Use of Laboratory Animals, Guide for the Care and Use of Agricultural Animals in Research and Teaching,* and the "European Convention for the Protection of Vertebrate Animals Used for Experimental and Other Scientific Purposes" (ETS 123; AAALAC "Rules of Accreditation", 2016). While there are questions about whether AAALAC programs create better conditions for animal care, the persistence of an accreditation body and committee seems to allay some public concerns about animal welfare in research settings (Goodman, Chandna and Borch, 2015).

BOX 2: Abbreviated US Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training

1. Transportation, care and use must conform to the Animal Welfare Act (7USC2131)
2. "Procedures involving animals should be designed and performed with due consideration of their relevance to human and animal health, the advancement of knowledge or the good of society"
3. Species and quantity of animals should be an appropriately selected minimum number to obtain valid results. Alternative methods—mathematical models and in vitro studies—are preferred to in vivo studies.
4. Use of techniques to avoid or minimize "discomfort, distress, and pain" are imperative.
5. Appropriate "sedation, analgesia, or anaesthesia" should be used when "more than momentary or slight pain or distress" or surgical or painful procedures are used.
6. "Animals that would otherwise suffer severe or chronic pain or distress that cannot be relieved should be painlessly killed at the end of the procedure or, if appropriate, during the procedure"
7. Living conditions should be species appropriate and supervised by a veterinarian.
8. Individuals involved in animal research should be properly trained and appropriately qualified.
9. Any exceptions to these principles must be made by an appropriate review committee and must meet the standards of principle 2.

Recombinant DNA (rDNA) Research

Enormous public and scientific concern about the use of recombinant nucleic acids (rDNA) in research was the primary driver for development of the newest form of research review board, Institutional Biosafety Committees (IBC) (Berg and Singer 1995). In fact, it was the published comments derived from a relatively small conference on nucleic acids—including the radical move to call for a voluntary moratorium on rDNA research—that moved concern about research on recombinant DNA into the public forum (Berg, Baltimore et al 1975; Singer and Soll 1973; Talbot 1980). While the promises of rDNA research, such as improved crops and medicines, were hailed as saviors of a world with dwindling capacity and resources, the possibility of unforeseeable effects stoked scientific and public fears of a man-made apocalypse (Race and Hammond 2008). Concerns of the scientific community included: 1) transparency of the research, 2) safety of the research (including researchers and infrastructure), and 3) control of research products (Ross et al 1996). The mechanisms of control imposed on the community of researchers working with synthetic nucleotides following the 1974/75 moratorium on rDNA research addressed the concerns through institution of risk-assessing review boards: Institutional Biosafety Committees (Berg et al 1974, 1975; Talbot 1983). As Jenkins summarizes,

> The role of the IBC is to ensure adequate containment of potentially hazardous biological agents; add a level of expert review and monitoring of potentially hazardous experiments; to inform the public about experimental plans that have a potential to be hazardous; and to provide a means of communication among researchers and healthcare providers about potentially hazardous protocols. The fundamental core of IBC review is the concept of risk assessment of work with biological materials (Jenkins, 2004, 16).

The conduct of rDNA research is governed through structures designed to minimize possible harms through appropriate risk review and mitigation strategies (Petrella 2014). Risk assessment includes evaluation of "virulence, pathenogenicity, infectious dose, environmental stability, route of spread, communicability, operations, quantity, availability of vaccine or treatment, and gene product effects such as toxicity, physiological activity, and allergenicity" (NIH 2013, quoted in Jenkins 16). To simplify review, agents (organisms and toxins) were sorted into four risk groups corresponding with differing levels of review; the greater the probability and magnitude of harm associated with a particular agent (e.g., risk group 4 research), the greater the scrutiny by the Institutional Biosafety Committee (Boreano 1984; WHO 2004).

Establishment of risk groups and standard control regulations per risk group guides investigators in their discussion of the risks of their projects but also guides review committees to understand when it may be necessary to elevate review of risky projects. The NIH Guidelines propose 6 levels of experiments requiring review, with those falling into level III-A, so called "major actions", being subject to the most extensive review.

The levels of review and examples of research meeting these levels are described in Box 3 "Levels of Review and Types of Research Covered by IBC Boards". Institutional Biosafety Committees are unusual in this set of three review board types as there was, until revisions to the NIH Guidelines in April 2019, a national level of review institutionalized in the Recombinant DNA Advisory Committee (RAC), which operates at the level of the NIH Directors' office (Kahn 2009). The unique feature of the RAC requires some explanation: until the most recent policy revisions, the 15 member RAC was responsible for conducting public hearings pertinent to research projects whose scientific, ethical, legal, and community concerns are deemed by local committee or by the NIH to warrant additional scrutiny. Subsequent to the April 2019 revisions, the RAC was re-envisioned as the NExTRAC (Novel and Exceptional Technology and Research Advisory Committee) with an expanded focus on "the transparent discussion of science, safety, and ethics" (NIH 2019).

Like IRBs, IBCs must be registered and must demonstrate their compliance with procedural and personnel requirements during their registration. IBCs register with the NIH Office of Science Policy, assuring the NIH OSP that they have "at least 5 members" with "appropriate recombinant and synthetic nucleic acid expertise collectively", "plant and animal experts, biosafety officer as appropriate" and "at least two members not affiliated with the institution" (Section IV-B-2-a of the NIH Guidelines). Ideally, other members are involved to bring in additional expertise in environmental and public health, law, physical containment (facilities) and laboratory technical personnel. The Biosafety Officer must be part of the IBC if the institution conducts large scale or "high containment" (BSL-3 or BSL-4) research (Choosewood and Wilson 2007).

BOX 3: Levels of Review and Types of Research Covered by IBC Boards

Level of review	Example of types of research covered	Relevant section(s) of the NIH Guidelines
IBC and NIH Director review and approval	Experiments involving deliberate transfer of a drug resistance trait to a microorganism when such resistance could compromise the ability to control the disease agent in humans, veterinary medicine, or agriculture	III-A
IBC approval and NIH OSP review for containment determinations	Experiment involving the cloning of toxin molecules with LD50 of less than 100 nanograms per kilogram of body weight	III-B
IBC and IRB approval and NIH review before research participant enrollment	Experiments involving the deliberate transfer of recombinant or synthetic nucleic acid molecules into a human research participant	III-C
IBC approval before initiation	Creating stable germline alterations of an animal's genome, or testing viable recombinant or synthetically modified microorganisms on whole animals, where BL-2 containment or greater is necessary	III-D
IBC notice at initiation	Creating stable germline alterations of rodents by introduction of recombinant or synthetic nucleic acid molecules when these experiments require only BL-1 containment	III-E
Exempt from the NIH Guidelines. IBC registration not required if experiment not covered by Sections III-A, III-B, or III-C	Purchase or transfer of transgenic rodents	III-F

IBC's remit may extend beyond rDNA research: they may also review the use of biological agents and toxins—select agents—that are regulated under the aegis of the Agricultural Bioterrorism Protection Act of 2002 and the Public Health Security and Bioterrorism Preparedness and Response Act of 2002 (Butler et al 2012). IBCs reviewing this type of research serve as part of the national security infrastructure by regulating "Dual Use Research" (National Institutes of Health 2012). Per the United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern (DURC), "DURC is life sciences research that, based on current understanding, can be reasonably anticipated to provide knowledge, information, products or technologies that could be directly misapplied to pose a significant threat with broad potential consequences to public health and safety, agricultural crops and other plants, animals, the environment, materiel, or national security" (National Institutes of Health 2012). DURC regulations stipulate that experiments which 'enhance the harmful consequences' of agents, 'disrupts immunity or the effectiveness of an immunization against the agent', confers resistance to useful prophylactics, increases conditions for successful dissemination of the agent, 'alters the host range' of the agent, 'enhances the susceptibility of a host population' or 'generates or reconstitutes an eradicated or extinct agent' are of sufficient concern to require the intervention of agencies. When planned research involves select agents or methods listed in the DURC policies, agencies are required to ensure a thorough review, develop a risk mitigation strategy, possibly consult with the National Science Advisory Board for Biosecurity, and report to this board and components of the Department of Homeland Security to ensure adequate monitoring of the material and methods of the research.

In this second part of this article, the argument is developed that IBCs may provide the best analogue to the structure and function of an AIRC, as I will describe on the following page.

INSTITUTIONAL ARRANGEMENT OF OVERSIGHT FOR ARTIFICIAL INTELLIGENCE RESEARCH

What can be learned from this review for the purpose of structuring an Artificial Intelligence Research Committee (AIRC)? The lengthy historical and institutional review of other review boards given above was done in the service of identifying through precedent some ideal characteristics for an artificial intelligence review committee (see Box 4, "Precedents for an AIRC Committee").

BOX 4: Precedents for an AIRC Committee Adopted from IRBs, IACUCs, and IBCs

From the history of human subjects review concerns and Institutional Review Boards, an AIRC would benefit from:

- Establishment of a Belmont Report like document that outlines easily understood principles for review,
- Establishment of an assurance, not merely registration, system for AIRCs, and
- Establishment of a public database of research projects modelled on clinical trials registries.

From the history of animal care and use (IACUC) committees, well designed AIRCs might borrow:

- Explicit and required incorporation of experts directly involved in the life cycle of AI systems (aka Robot veterinarians), and
- As it becomes available when programmed according to the principles established through an AI Belmont Report, review of AI research by AI.

From the organization of institutional biosafety committees, AIRCs could benefit from inclusion of:

- Direct paths for elevation of potentially risky research to review at the federal level (an F-AIRC that functions like the NExTRAC),
- Establishment of clear categories of research experiments requiring review,
- Statement of clear "levels of AI safety" commensurate with the 4 biosafety levels, and
- Incorporation of national-security concerns for research with defense implications into the remit of review.

Purpose of the Committee

An Artificial Intelligence Review Committee makes decisions about the risks to human well-being, animals, the biome, and the national security infrastructure that stem may arise from artificial intelligence research, to include machine learning, big-data, and neural network research. The goal of an AIRC is not to implement technological feasibility analysis or to critique research methods, but to answer narrow questions about the risk of harm and types of harm arising from specific projects. An AIRC should also not engage in adjudication of disputes about the overall value of AI research or seek to oversee basic research that is only provisionally related to AI in a hypothetical future.

Which Agency or Department Leads?

Perhaps the most significant organizational design choice for an AIRC system is selection of agency to oversee AIRCs, issue relevant regulations, and superintend assurances. Constituting a Federal Robotics Commission is a choice with historical precedent, but whether this is an Independent Regulatory Commission, part of an existing organization such as Office of Science and Technology Policy in the Executive Office of the President, or unit of the Office of Research Integrity, is an institutional design question that would also need to be solved through acts of Congressional authority, like those seen in the Animal Welfare Act (Calo 2017, 402; Tutt 2017). A federal level AIRC, serving in a role like the newly constituted NExTRAC, would need to be explicitly authorized by Congress to engage in regular review of projects and to integrate with appropriate offices, such as the Office of Human Research Protections within the Department of Health and Human Services. In addition, a federal level AIRC should have a relationship with relevant experts within the Defense infrastructure and with the national laboratories whose research projects or facilities may be essential sites for AI research.

Learning from IRBs

Institutional Review Boards are the most visible and most maligned of the review committee structures (Gunsalus et al 2007). IRBs also have a strong basis of clear ethical principles, captured in the *Belmont Report*, for review of research from which to draw justifications for their reviews. A successful AIRC ought to cultivate the same, clear, principles-based review structure in order to ensure ethical cohesion, even in the absence of procedural symmetry achieved through a coordinated assurances system. However, establishment of using an assurance system and registries for projects, modeled on the registry for IRBs and for clinical trials, will help to increase public trust in research and research review. Establishing public trust in local boards is a problem that IRBs confronted and then resolved through the Federalwide Assurance system. Although the FWA system does not guarantee that all research is reviewed in precisely the same way, this system does create a set of essential organizational requirements for effective review regardless of local conditions that can be verified by the public. In a similar vein, a projects registry ensures that the public has a single source for substantive information about projects. A projects (e.g., clinical trials) registry, provides the public with essential knowledge about projects including location, duration and funding sources (Laine et al 2007; Macrae 2007). The mission and model of the WHO clinical trials registry (International Clinical Trials Registry Platform) offers guidance for trial database construction: these databases should document as “complete [a] view of research” to assist in health care decision making and to “strengthen the validity and value of the scientific evidence base” (WHO). While registries may pose

challenges to the practices of keeping elements of algorithms or data protected as trade secrets, the history of protecting proprietary interests of pharmaceutical companies' projects registered on publicly accessible registries provides a precedent for overcoming these concerns.

Lessons from Animal Care and Use Committees

As animal welfare committees evolved over time, the number and relative power of veterinarians on the committees has grown, whether as an act of will by individual committees or as a response to regulations. Within the context of the establishment of an AIRC, it is necessary that individuals with deep, expert, knowledge of artificial general intelligence and autonomous system programming be included as players with substantive, not merely procedural, power and consequence. Over time, there has been a call for inclusion of anti-vivisection advocates or members of animal welfare advocacy groups on animal care and use committees. The purpose of their inclusion is to represent the interests of the animals. As AI evolves to become more like a fully autonomous system—when, perhaps, these pass the Arkin test (1998, 2010)—it would seem ethically incumbent upon committees to include AIs into the system as a representative of the affected community. Until AI becomes sufficiently intelligent or autonomous, the inclusion of individuals who "speak for the systems" just as veterinarians or animal-rights activists speak for the animals, seems ethically good, if not eventually legally right.

Lessons from IBCs

Biosafety and Select Agents

As the latest committee to be instituted, and as the review board structure examining scientific research with the greatest degree of uncertainty around the relevant science, the IBC represents what might be the best analogue for an appropriate AIRC system. The following four components might be useful components for inclusion in the overall AIRC structure: federal level committee, risk-adjusted review, management of dual use concerns, and designation of an institutional official.

F-AIRC

For both IRBs and IACUCs, the local institution is the arbiter of research approval or, in some cases, advancing review to a higher level. For Institutional Biosafety Committees, the RAC (recently reconstituted as the NExTRAC) operating in the National Institutes of Health provides a federal level review resource for projects with considerable risk. Given the high degree of uncertainty regarding replication of inserted sequences, transformation of organisms through deletion or other mutations, and production of uncontrolled and undesirable traits as a product of research, the NExTRAC serves as an additional check on this uncertainty. Publication of NExTRAC hearings and decisions in the *Federal Register* increases public transparency, thus reducing fear surrounding this necessary uncertainty. With similar degrees of uncertainty surrounding true risk of harm from AI research, the possibility of replication of undesirable or uncontrollable "traits", or potential for unintentional transformation of related systems, a federal level AIRC seems a warranted part of the AI review landscape (Bostrom 2003, 2012). As was described above, the federal level committee does not review all rDNA and associated research: federal level review is reserved for research that requires a major review action and is advanced to the federal level by a duly constituted board who determines the project meets specific guidelines regarding the risks to humans, animals, economies, and other systems.

AI-Safety Risk-Group and Containment Levels

As researchers interested in IRB conduct have noted, the lack of specificity in determination of research risk can lead to vastly different review demands (Boreano 1984; Green 2010). To mitigate against this hazard for the AI community, particularly given the disparities in distribution of AI researchers across states and nations, more rigorous categories of research risk, such as those designated for IBCs could be adopted (DeAngelis et al 2004; Dyrbye et al 2007).

Development of AI risk groups and AI "containment" manuals would be an ideal way to streamline AI research review and to increase the transparency of review committee decisions. AI risk group levels might follow a similar logic to biological risk levels wherein severity of harm, availability of preventive mechanisms, and availability of therapeutic materials determines the risk group level (Biosafety in Microbiological and Biomedical Laboratories manual; see also Choosewood and Wilson 2007). The following boxes outline possible extrapolations of these risk groups to Artificial Intelligence Risk Groups and, considering the relationship of AI to autonomous systems development, Autonomous Systems Risk Groups (the following works are examples of those consulted to develop these categories: Anderson et al 2014; Arkin and Balch 1997; Dougherty and Giardina 1988; Littman 2014; Lucas 2013; Maes 1990; Pennachan 2007; Steels 1993; Varela and Bourgine 1992).

BOX 5: Artificial Intelligence Risk Group Categories with Examples

Artificial Intelligence Risk Groups 1–4

- AI-RG-1: no threats to human or animal decision-making capacity or welfare are associated with this program
  - » Gaming algorithms that predict player performance and adjust gaming environments
- AI-RG-2: limited interference with human decision-making capacity and welfare or animal health and safety could arise if preventative steps or corrective counter-measures are not taken in a timely manner
  - » Autocorrect algorithms that reduce human capacity for correct and grammatical communication
  - » Physician diagnostic algorithms the use of which reduce diagnostic ability without the systems
- AI-RG-3: significant interference with human decision-making capacity, welfare, and self concept or with animal health, safety, and species integrity may arise from use of this system. Development of these systems should be done in tandem with development of corrective counter-measures for misuse or loss of control
  - » Emotionally imprinting AI systems, such as those envisioned in the film "AI"
- AI-RG-4: irreparable interference with human capacities and self-concept or animal welfare and species integrity for which corrective countermeasures for misuse or loss of control represent distant research and development horizons
  - » Artificial superintelligent systems that can, when permitted to self-replicate, exceed the boundaries of human or animal control (e.g., "grey goo")

BOX 6: Autonomous Systems Risk Group Categories with Examples

Autonomous Systems Risk Groups 1–4

- AS-RG-1: no threats to human or animal decision-making capacity or welfare are associated with deployment of this system
  - » Self-guided vacuum units (e.g., Roomba)
- AS-RG-2: use of the system may lead to decisional or functional dependency for some users, but preventive or corrective techniques are readily accessible and well developed
  - » NHTSA Level 3 autonomous vehicles
- AS-RG-3: use of the system is likely to lead to decisional or functional dependency and possible loss of self-concept for some users, but preventive or corrective techniques are available, even if in an early stage of development. Development of these systems should be done in tandem with development of preventive or therapeutic systems.
  - » NHTSA Level 4 autonomous vehicles
- AS-RG-4: irreparable interference with human decisional or functional abilities, including total dependence for which preventive or therapeutic measures are unknown or represent distance research and development horizons.
  - » Fully autonomous, humanoid, robots

Those instances where a research project presents serious concerns, perhaps level 3 or level 4 concerns should trigger a higher level of review, such as a federal level AIRC review. See Box 7 for an extrapolation AI safety.

What about Dual Use Research in AI?

Review of select agent projects , those biological or toxicological research projects that use one or more of the agents identified as having potential to be weaponized, now rests within the remit of many IBCs. According to the pertinent regulations, implemented by the HHS and USDA, "select agents and toxins are a subset of biological agents and toxins ... determined to have the potential to pose a severe threat to public health and safety, to animal or plant health, or to animal or plant products" (42 USC 262a and 7 USC 8401). These agents fall under the definition of "Dual Use Research of Concern" (DURC), which has distinct requirements for incorporation of risk mitigation measures to ensure that the research follows the oversight reporting requirements of the sponsoring agencies, the advice of the National Science Advisory Board for Biosecurity, and classification requirements under the National Security Decision Directive 189.

The tools of artificial intelligence could also be used to threaten public health and safety. Explicit development of systems with artificial intelligence components designed to surveille or constrain individuals, even if developed initially for peaceful or domestic law enforcement, will trigger DURC concerns. Development of tools that could be imparted to existing weapons systems to enhance their capabilities or developed into wearables or implantable tools that could alter the cognitive

BOX 7: Adaptation of the IBC-RAC Categories of Research for an AIRC/ F-AIRC System

Level of review	Example of types of research covered
AIRC and F-AIRC review	Artificial Intelligence Risk Group (AI-RG) 3 & 4 Autonomous Systems Risk Group (AS-RG) 3 & 4 Deployment of AI or AS without corrigibility or multiply redundant failsafe "kill switch" mechanisms
AIRC and F-AIRC review for advice including containment or therapeutic amelioration determinations, possible IRB or IBC review	AI-RG 3 and AS-RG 3: Deployment of an artificial intelligence system with self-replicating or self-correcting ability in a minimally controlled, networked, environment OR with a human or animal population outside of direct research settings (e.g., robot imprinting in a family dynamic setting) Testing of an autonomous system with a representative sample of human subjects in an uncontrolled setting (naturalistic driving studies of a novel computer vision system) Testing of systems of multiply redundant failsafe or "kill switch" mechanisms in mature AI or AS systems
AIRC and IRB or IACUC approval and F-AIRC review for advice before research participant enrollment or animal-system interaction permitted	AI-RG 2 & 3 and AS-RG 3 Pilot studies of proof of concept autonomous system consumer goods in a controlled setting with healthy volunteers Pilot studies of proof of concept for adaptive and self-correcting learning algorithms in a controlled setting with healthy and monitored volunteers
AIRC approval before initiation	AI-RG 2 and AS-RG 2 Proof of concept tests for new autonomous system based consumer goods with no direct human or animal interaction Proof of concept tests for learning algorithms to adapt automated testing environment to human or animal stress indicators with no interaction
AIRC notification before initiation	AI-RG 1 and AS-RG 1 Reverse engineering existing systems in a course or laboratory setting where outcomes of reversal are uncertain
Exempt from AIRC review	Teaching demonstrations of AI or AS on contained or non-networked systems Purchase of commercially available AI or AS for the purpose of researching its characteristics or use in teaching

readiness of soldiers will also trigger review of AI under the terms of DURC. Further, the export of this technology to others will likely trigger review of the research and its products under the aegis of import-export controls" (Leung, Fischer and Dafoe 2019). Review of DURC, weaponizeable or defense systems, ought to be conducted by an AIRC, though preferentially this should occur at the federal AIRC level.

Minimum Procedural Conditions for Establishment of an AIRC

Thus far, this article developed analogies for design of decision-tools and organization for artificial intelligence review committees. An ideal AIRC committee would fulfill the roles and responsibilities, borrowed from the procedural requirements placed upon the boards described above and described in Box 7 below. But, the basics of committee administration, such as committee meeting coordination techniques (e.g., electronic document management systems), member training programs, protocol review techniques, and funding models (e.g., institutional research funds levied through "facilities and administration costs" or through protocol review fees) should be investigated based upon the literature surrounding these other review committees to determine optimal initial designs (Sugarman 2000).

BOX 8: Procedural Requirements for an Artificial Intelligence Review Committee

Drawing upon the organizational requirements from other review boards, an AIRC should have:

- Registration and submission of assurances to a designated governmental organization
- Constitution at the local level at a similar level of power and function as other review committees
  - » Cross-collaborations established with IRB, IACUC, and IBC
- Guidance of an Institutional Official (IO) designated to accept the responsibilities and obligations commensurate with such a review board
- A professional administrative staff that should keep appropriate minutes, protocol and project review calendars, and track amendments
- Membership of at least 7 human members that represent artificial intelligence design and programming expertise, including at least one non-affiliated community member
- One member to speak "for" AI as its capabilities increase

  Have a well-designed and appropriate protocol review form that includes areas for description of at least:
  - » PI, Co-PI and research team information and expertise
  - » Description of research project including description of systems produced, participants interacting with the systems (if any), expected risks and benefits to research team and to the appropriate community, project duration, project implementation locations, network accessibility and cybersecurity measures, and any industry attachments or conflicts of interest from the researchers
  - » Research results dissemination plan
  - » Dual-Use declaration or Import/Export Regulatory Assurance

CONCLUDING CONCERNS

Review boards are organs of social engineering that operate in the context of the projects they govern and the history of review board regulation and functions. To accomplish the goal of engineering the social context of artificial intelligence research, the experiences of existing review boards' establishment and conduct must be consulted in order to mitigate known hazards (Walter and Klein 2003). These known hazards include establishment of a rigid hierarchy and hardened-fast review procedures argued to lead to overly bureaucratic review mechanisms which limit researcher freedom and heap costs of time and money onto grants and contracts (Butzow 1995; Emanuel et al, 2004; Gunsalus et al 2007; Jacobson, Gewortz and Haydon 2007; Ozdemir 2009; White 2007). Rigid regulations governing review boards themselves have also led to well-meaning institutions running afoul of oversight bodies when they let compliance with one rule or another "slip" (Marshall 1999). Attempts to side step regulatory problems led historically to other hazards such as politicization, stacking the deck of committee membership leading to narrow thinking, capture by powerful interests, and concentration of expertise assets at the board governance level not the review board itself. Failure to give the committee sufficient resources, including investigational and enforcement authority, or to set up appropriate appeal mechanisms, is known to lead to problems of investigator frustration, purposeful scofflaw behavior by investigators, and charges of unfairness in review (Bankert and Amdur 2006; Dyrbye et al 2007; Fiske 2009; Selgelid 2009; Singer 2001). Finally, failure to situate the power of the review board infrastructure within a sufficiently powerful agency or group of agencies could lead to such a board having too little policing powers to support their normative force.

In the concluding paragraphs of this piece, two final issues of concern are broached with the intent of spurring additional consideration and recommendations for a future Artificial Intelligence Review Committee. These are: 1) with whom does the "buck stop" if AI research produces harms? And 2) where should global governors be in this process?

Who "wears the orange jumpsuit"[2] if AI research goes awry?

Identifying the site for the highest review authority for AI research is a design problem for serious consideration. Limiting the context of the discussion to the US, Calo calls for a Federal Robotics Commission, but unless such a commission is situated at the level similar to the existing NIH level RAC, it seems infeasible that this committee could function as the number and range of projects may overwhelm its capacity. Siting all review authority in one body at this level does offer a simple solution to solve the problem of accountability. Realistically, however, initial review will be localized as it is in the IBC, IRB and IACUC, which raises the perennial matter of jurisdictional intersections that complicate the determination of when an Institutional Official could be fined or jailed for approving research or failing to elevate its approval to another level of oversight (Sugarman 2000). Finally, if it is an autonomous system or a learning system which causes harm, then whether the researcher, the Institutional Official, or the review boards are "really" responsible is a question that will complicate efforts to establish liability for the harms (Cerka, Grigiene, and Sirbikyte 2015, Hallevy 2010, Vladeck 2014). Future legal and ethical research must grapple with effective review of learning systems soon.

---

2 This phrase is credited to Dr. Michael Buckley manager of the IRB in the Vice President for Research Office at Texas A&M University in 2002. An orange jumpsuit is a reference to prisoners' uniforms.

Where should global governors be in this process?

Autonomous systems and artificial intelligence research is not a nationally bound research endeavor. Many AI projects are internationally collaborative and most cross boundaries of institutions, such as government-funded academic research that uses industry data. These cross-cutting collaborations raise further issues for building review-boards with authority, such as determining whether global governing bodies, such as ISO or IEEE, are the appropriate organizations for determining the standards for evaluation of ethical implications of AI research. These two bodies, with IEEE being foremost as a professional organization dedicating itself to these concerns presently, are relevant global governors whose role as standard setting bodies or even the optimal hosts for supranational review committee organizations should be addressed in near-term future research (Bruce and Biersteker 2002; Murphy and Yates, 2009; Roco 2008). Research and workshops to address these two issues and the myriad others alluded to, but not confronted head on in the above paragraphs, ought to be pursued earnestly.

It is inevitable that the adjustment of the AI and AS research enterprise to a prospective review structure synonymous with that of human subjects, animal use, and rDNA research will be an uncomfortable one. However, the continuation of general calls for review should be eclipsed by hardened examination of review board design and implementation of review mechanisms. This should be done sooner rather than later, before the calls for review of AI research become jeremiads lamenting what should have been done.

WORKS CITED

- AAALAC (Association for Assessment and Accreditation of Laboratory Animal Care). "About AAALAC". (2016). Available at: http://aaalac.org/about/index.cfm
- Anderson, James M., Kalra Nidhi, Karlyn D. Stanley, Paul Sorensen, Constantine Samaras, and Oluwatobi A. Oluwatola. *Autonomous vehicle technology: A guide for policymakers*. Rand Corporation, 2014.
- Arkin, Ronald C. *Behavior-based robotics*. MIT press, 1998.
- Arkin, Ronald C. "The case for ethical autonomy in unmanned systems." *Journal of Military Ethics* 9, no. 4 (2010): 332-341.
- Arkin, Ronald C., and Tucker Balch. "AuRA: Principles and practice in review." *Journal of Experimental & Theoretical Artificial Intelligence* 9, no. 2-3 (1997): 175-189.
- Bankert, Elizabeth A., and Robert J. Amdur. *Institutional review board: Management and function*. Jones & Bartlett Learning, 2006.
- Beecher, Henry K. "Consent in clinical experimentation: myth and reality." *JAMA* 195, no. 1 (1966): 34-35.
- Beecher, Henry K. "Ethics and clinical research." In *Biomedical ethics and the law*, pp. 193-205. Springer US, 1976.
- Berg, Paul, David Baltimore, Herbert W. Boyer, Stanley N. Cohen, Ronald W. Davis, David S. Hogness, Daniel Nathans et al. "Potential biohazards of recombinant DNA molecules." *Science* 185, no. 4148 (1974): 303.
- Berg, Paul, David Baltimore, Sydney Brenner, Richard O. Roblin, and Maxine F. Singer. "Summary statement of the Asilomar conference on recombinant DNA molecules." *Proceedings of the National Academy of Sciences* 72, no. 6 (1975): 1981-1984.
- Berg, P, and M F Singer. "The recombinant DNA controversy: twenty years later." Proceedings of the National Academy of Sciences of the United States of America vol. 92, 20 (1995): 9011-3. doi:10.1073/pnas.92.20.9011
- Bereano, Philip L. "Institutional biosafety committees and the inadequacies of risk regulation." *Science, Technology, & Human Values* 9, no. 4 (1984): 16-34.
- Bostrom, Nick. "Ethical issues in advanced artificial intelligence." *Science Fiction and Philosophy: From Time Travel to Superintelligence* (2003): 277-284.
- Bostrom, Nick. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents." *Minds and Machines* 22, no. 2 (2012): 71-85.
- Brandt, Allan M. "Racism and research: the case of the Tuskegee Syphilis Study." *Hastings Center Report* 8, no. 6 (1978): 21-29.
- Butler, R. Mark, Nancy D. Connell, Steven S. Morse, Mark Campbell, and Raymond C. Tait. "Strengthening the role of the IBC in the 21st century." *Ensuring National Biosecurity: Institutional Biosafety Committees* (2016): 217.
- Butzow, Mary. "Getting through the IRB: Covenant Medical Center Nursing Research Committee's experience." *Journal of Emergency Nursing* 21, no. 3 (1995): 262-263.
- Calo, Ryan. "Robotics and the Lessons of Cyberlaw." *Cal. L. Rev.* 103 (2015): 513.
- Calo, Ryan. "Robots and privacy." *ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS,* Patrick Lin, George Bekey, and Keith Abney, eds., Cambridge: MIT Press (2010).
- Calo, Ryan. "Bring on the Robocrats." *Scientific American* 311, no. 6 (2014): 12-12.
- Calvo, Rafael and Dorian Peters. "AI Surveillance studies need ethics review". *Nature* 557, no 31 (2018): 1. doi: 10.1038/d41586-018-05015-1
- Carbone, Larry. *What animals want: Expertise and advocacy in laboratory animal welfare policy*. Oxford University Press, USA, 2004.
- Cassell, Eric J. "The principles of the Belmont report revisited: How have respect for persons, beneficence, and justice been applied to clinical medicine?." *Hastings Center Report* 30, no. 4 (2000): 12-21.
- Čerka, Paulius, Jurgita Grigienė, and Gintarė Sirbikytė. "Liability for damages caused by artificial intelligence." *Computer Law & Security Review* 31.3 (2015): 376-389.
- Chosewood, L. Casey, and Deborah E. Wilson. *Biosafety in microbiological and biomedical laboratories*. Diane Publishing, 2007.
- Clark, J. D., R. L. Baldwin, K. A. Bayne, M. J. Brown, G. F. Gebhart, J. C. Gonder, J. K. Gwathmey et al. "Guide for the care and use of laboratory animals." *Washington, DC: Institute of Laboratory Animal Resources, National Research Council* 125 (1996).
- Committee on Publication Ethics. "Code of Conduct and Best Practice Guidelines for Journal Editors". (2011). Available at: http://publicationethics.org/files/Code%20of%20Conduct_2.pdf
- Cottingham, John. "'A Brute to the Brutes?': Descartes' Treatment of Animals." *Philosohy* 53, no. 206 (1978): 551-559.
- Crichton, Michael. *The Andromeda strain*. Random House, 1993.
- De Angelis, Catherine, Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, Sheldon Kotzin et al. "Clinical trial registration: a statement from the International Committee of Medical Journal Editors." *New England Journal of Medicine* 351, no. 12 (2004): 1250-1251.
- Demers, Gilles, Gilly Griffin, Guy De Vroey, Joseph R. Haywood, Joanne Zurlo, and Marie Bédard. "Harmonization of animal care and use guidance. *"Science* 312, no. 5774 (2006): 700-701.

- Department of Health and Human Services. "Protection of Human Subjects" Revised January 15, 2009. Available at: http://www.hhs.gov/ohrp/regulations-and- policy/regulations/45-cfr-46/index.html
- Dougherty, Edward R. Edward R., and Charles R. Giardina. *Mathematical methods for artificial intelligence and autonomous systems*. No. 04; Q336, D6. 1988.
- Dyrbye, Liselotte N., Matthew R. Thomas, Alex J. Mechaber, Anne Eacker, William Harper, F. Stanford Massie Jr, David V. Power, and Tait D. Shanafelt. "Medical education research and IRB review: an analysis and comparison of the IRB review process at six institutions." *Academic Medicine* 82, no. 7 (2007): 654-660.
- Eden, Amnon H., James H. Moor, Johnny H. Soraker, and Eric Steinhart. *Singularity hypotheses: A scientific and philosophical assessment*. Springer Science & Business Media, 2013.
- Edgar, Harold, and David J. Rothman. "The institutional review board and beyond: future challenges to the ethics of human experimentation." *The Milbank Quarterly* (1995): 489- 506.
- Emanuel, Ezekiel J., Anne Wood, Alan Fleischman, Angela Bowen, Kenneth A. Getz, Christine Grady, Carol Levine et al. "Oversight of human participants research: identifying problems to evaluate reform proposals. *"Annals of internal medicine* 141, no. 4 (2004): 282-291.
- Fiske, Susan T. "Institutional review boards: From bane to benefit. *"Perspectives on Psychological Science* 4, no. 1 (2009): 30-31.
- Frumkin, Peter, and Joseph Galaskiewicz. "Institutional isomorphism and public sector organizations." *Journal of public administration research and theory* 14, no. 3 (2004): 283-307.
- Giles, Jim. "Nanotech takes small step towards burying 'grey goo'." *Nature*429, no. 6992 (2004): 591-591.
- Goertzel, B., and P. Wang. "A foundational architecture for artificial general intelligence." *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* 6 (2007): 36.
- Goodman, Justin R., Alka Chandna, and Casey Borch. "Does accreditation by the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC) ensure greater compliance with animal welfare laws?." *Journal of Applied Animal Welfare Science* 18, no. 1 (2015): 82-91.za
- Grady, Christine. "Do IRBs protect human research participants?." *JAMA*304, no. 10 (2010): 1122-1123.
- Green, Lee A., Julie C. Lowery, Christine P. Kowalski, and Leon Wyszewianski. "Impact of institutional review board practice variation on observational health services research." *Health services research* 41, no. 1 (2006): 214-230.
- Gunsalus, C. Kristina, Edward M. Bruner, Nicholas C. Burbules, Leon Dash, Matthew Finkin, Joseph P. Goldberg, William T. Greenough et al. "The Illinois white paper improving the system for protecting human subjects: Counteracting IRB "Mission creep"." *Qualitative Inquiry* 13, no. 5 (2007): 617-649.
- Hackney, Raymond W., Theodore A. Myatt, Kathleen M. Gilbert, Rebecca R. Caruso, and Susanne L. Simon. "Current trends in institutional biosafety committee practices." *Applied Biosafety* 17, no. 1 (2012): 11-18.
- Hall, Rodney Bruce, and Thomas J. Biersteker. *The emergence of private authority in global governance*. Vol. 85. Cambridge University Press, 2002.
- Hallevy, Gabriel. "The criminal liability of artificial intelligence entities-from science fiction to legal social control." *Akron Intell. Prop. J.* 4 (2010): 171.
- Hartzog, Woodrow. "Unfair and Deceptive Robots." *Maryland Law Review* 74, no. 785 (2015).
- Hernández-Orallo, José, and David L. Dowe. "Measuring universal intelligence: Towards an anytime intelligence test." *Artificial Intelligence* 174, no. 18 (2010): 1508-1539.
- Interagency Research Animal Committee. "US government principles for the utilization and care of vertebrate animals used in testing, research and training." *Available at: grants.nih.gov/grants/olaw/references/phspol.htm#USGovPrinciples. Accessed Dec* 13 (2010).
- Jacobson, Nora, Rebecca Gewurtz, and Emma Haydon. "Ethical review of interpretive research: Problems and solutions." *IRB: Ethics & Human Research* 29, no. 5 (2007): 1-8.
- Jenkins, Christopher. "Trends in united states biological materials oversight and institutional biosafety committees." *Journal of Research Administration* 45, no. 1 (2014): 11-47.
- Jennings, Nicholas R. "On agent-based software engineering." *Artificial intelligence* 117, no. 2 (2000): 277-296.
- Kahn, Jeffrey P. "Commentary: Who's Afraid of the RAC? Lessons from the Oversight of Controversial Science." *The Journal of Law, Medicine & Ethics*37, no. 4 (2009): 685- 687.
- Katz, Jack. "Toward a natural history of ethical censorship." *Law & Society Review* 41, no. 4 (2007): 797-810.
- Kilkenny, Carol, William Browne, Innes C. Cuthill, Michael Emerson, and Douglas G. Altman. Animal research: reporting in vivo experiments: the ARRIVE guidelines." *British journal of pharmacology* 160, no. 7 (2010): 1577-1579.
- Laine, Christine, Richard Horton, Catherine D. DeAngelis, Jeffrey M. Drazen, Frank A. Frizelle, Fiona Godlee, Charlotte Haug et al. "Clinical trial registration—looking back and moving ahead." *New England Journal of Medicine* 356, no. 26 (2007): 2734-2736.

- Leung, Jade, Sophie-Charlotte Fischer, and Allan Dafoe. "Export controls in the age of AI". *War on the Rocks*. August 28, 2019. Available at: https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/
- Litman, Todd. "Autonomous Vehicle Implementation Predictions." *Victoria Transport Policy Institute* 28 (2014).
- Lucas Jr, George R. "Legal and Ethical Precepts Governing Emerging Military Technologies: Research and Use." *Utah L. Rev.* (2013): 1271.
- Macoubrie, Jane. "Public perceptions about nanotechnology: Risks, benefits and trust." *Journal of Nanoparticle Research* 6, no. 4 (2004): 395-405.
- Macrae, Duncan J. "The Council for International Organizations and Medical Sciences (CIOMS) guidelines on ethics of clinical trials." *Proceedings of the American thoracic society* 4, no. 2 (2007): 176-179.
- Maes, Pattie. "Situated agents can have goals." *Robotics and autonomous systems* 6, no. 1 (1990): 49-70.
- Marchant, Gary E., and Wendell Wallach. "Coordinating Technology Governance." *Issues in Science and Technology* 31, no. 4 (2015): 43.
- Marshall, Eliot. "Shutdown of research at Duke sends a message." *Science* 284, no. 5418 (1999): 1246-1246.
- Mendelson III, Joseph. "Should Animals Have Standing: A Review of Standing under the Animal Welfare Act." *BC Envtl. Aff. L. Rev.* 24 (1996): 795.
- Miller, Seumas, and Michael J. Selgelid. "Ethical and philosophical consideration of the dual-use dilemma in the biological sciences." *Science and engineering ethics* 13, no. 4 (2007): 523-580.
- Murphy, Craig N., and JoAnne Yates. *The International Organization for Standardization (ISO): global governance through voluntary consensus*. Routledge, 2009.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, and Kenneth John Pres Ryan. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research-the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*. US Government Printing Office, 1978.
- National Institutes of Health. "NIH Organizational Chart". (no date). Available at: https://interodeo.od.nih.gov/nihchart/docs/NIH_Org_Chart.pdf
- Newgard, Lewis. "The paradox of human subjects protection in research: Some thoughts on and experiences with the federalwide assurance program." *Acad Emerg Med* 9, no. 12 (2002).
- Office of Human Research Protections (OHRP). "Registering an IRB and Obtaining an FWA". (2016). Available at: http://www.hhs.gov/ohrp/register-irbs-and-obtain- fwas/steps/
- Office of Human Research Protections (OHRP). "Federal Policy for the Protection of Human Subjects". (2016) Available at: http://www.hhs.gov/ohrp/regulations-and- policy/regulations/common-rule/
- Office of Human Research Protections (OHRP). "The Belmont Report". (2016). Available at: http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.
- Ozdemir, Vural. "What to do when the risk environment is rapidly shifting and heterogeneous? Anticipatory governance and real-time assessment of social risks in multiply marginalized populations can prevent IRB mission creep, ethical inflation or underestimation of risks." *The American Journal of Bioethics* 9, no. 11 (2009): 65-68.
- Pennachin, Cassio, and Ben Goertzel. "Contemporary approaches to artificial general intelligence." In *Artificial general intelligence*, pp. 1-30. Springer Berlin Heidelberg, 2007.
- Petrella, Brenda L. "Biosafety Oversight and Compliance: What do you Mean, I have to Fill Out Another Form?!." *Current protocols in microbiology*(2014): 1A-5.
- Pray, Carl E., Bharat Ramaswami, Jikun Huang, Ruifa Hu, Prajakta Bengali, and Huazhu Zhang. "Costs and enforcement of biosafety regulations in India and China." *International Journal of Technology and Globalisation* 2, no. 1-2 (2006): 137-157.
- Race, Margaret S., and Edward Hammond. "An evaluation of the role and effectiveness of institutional biosafety committees in providing oversight and security at biocontainment laboratories." *Biosecurity and bioterrorism: biodefense strategy, practice, and science* 6, no. 1 (2008): 19-35.
- Resnik, D. B. (2009). *Playing politics with science: Balancing scientific independence and government oversight*. Oxford University Press on Demand.
- Roco, Mihail C. "Possibilities for global governance of converging technologies." *Journal of Nanoparticle Research* 10, no. 1 (2008): 11-29.
- Rodrigues, Rowena, Clare Shelley-Egan, Marlou Bijlsma, and Thamar Zijlstra. "Ethics Assessment and Guidance in Different Types of Organisations". SATORI Annex 3.i. Available at: http://satoriproject.eu/media/3.i-Standards-certification-and-accr-orgs.pdf
- Rollin, Bernard E. "The regulation of animal research and the emergence of animal ethics: a conceptual history." *Theoretical medicine and bioethics* 27, no. 4 (2006): 285- 304.Res
- Ross, Gail, Robert Erickson, Debra Knorr, Arno G. Motulsky, Robertson Parkman, Jude Samulski, Stephen E. Straus, and Brian R. Smith. "Gene therapy in the United States: a five-year status report." *Human gene therapy*7, no. 14 (1996): 1781-1790.

- Sandler, Ronalde, John Basl and Steven Tiell. "Building Data and AI Ethics Committees". Available at: https://cssh.northeastern.edu/informationethics/wp-content/uploads/sites/51/2019/08/811330-AI-Data-Ethics-Committee-Report_V10.0.pdf#_ga=2.47934759.1473589072.1567159258-1792561510.1567159258?mod=article_inline
- Selgelid, Michael J. "Governance of dual-use research: an ethical dilemma." *Bulletin of the World Health Organization* 87, no. 9 (2009): 720-723.
- Selgelid, Michael J. "Dual-use research codes of conduct: Lessons from the life sciences." *Nanoethics* 3, no. 3 (2009): 175-183.
- Silverman, Jerald, Mark A. Suckow, and Sreekant Murthy, eds. *The IACUC handbook*. CRC Press, 2014.
- Simes, Robert J. "Publication bias: the case for an international registry of clinical trials." *Journal of clinical oncology* 4, no. 10 (1986): 1529-1541.
- Singer, Maxine, and Dieter Soll. "Guidelines for DNA hybrid molecules. *"Science* 181, no. 4105 (1973): 1114.
- Singer, Maxine. "Commentary: What Did the Asilomar Exercise Accomplish, What Did it Leave Undone?." *Perspectives in biology and medicine* 44, no. 2 (2001): 186-191.
- Statt, Nick. "Google dissolves AI ethics board just one week after forming it". *The Verge*. April 4, 2019. Available at: https://www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation
- Steels, Luc. "The artificial life roots of artificial intelligence." *Artificial life* 1, no. 1_2 (1993): 75-110.
- Stone, Zara. "The Artificial Intelligence Ethics Committee". *Forbes*. June 11, 2018. Available at: https://www.forbes.com/sites/zarastone/2018/06/11/the-artificial-intelligence-ethics-committee/#4f178a461637
- Sugarman, Jeremy. "The role of institutional support in protecting human research subjects." *Academic Medicine* 75, no. 7 (2000): 687-692.
- Talbot, Bernard. "Introduction to recombinant DNA research, development and evolution of the NIH guidelines, and proposed legislation." *U. Tol. L. Rev.*12 (1980): 804.
- Talbot, B. "Development of the National Institutes of Health Guidelines for Recombinant DNA Research." Public health reports (Washington, D.C.: 1974) vol. 98, 4 (1983): 361-8.
- The IEEE Global Initiative. "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems" First Edition (2019). Available at: https://ethicsinaction.ieee.org/#read.
- Thomas, Stephen B., and Sandra Crouse Quinn. "The Tuskegee Syphilis Study, 1932 to 1972: implications for HIV education and AIDS risk education programs in the black community." *American journal of public health* 81, no. 11 (1991): 1498-1505.
- Todd. Deborah. "Microsoft Reconsidering AI Ethics Review Plan". *Forbes*. June 24, 2019. Available at: https://www.forbes.com/sites/deborahtodd/2019/06/24/microsoft-reconsidering-ai-ethics-review-plan/#526913e97c89
- Tutt, Andrew. "An FDA for algorithms". *Administrative Law Review* 69(1): 81-123.
- USDA, United States Department of Agriculture. "Animal Welfare Act". (no date). Available at: https://www.nal.usda.gov/awic/animal-welfare-act.
- Varela, Francisco J., and Paul Bourgine. *Toward a practice of autonomous systems: Proceedings of the First European Conference on Artificial Life*. MIT Press, 1992.
- Vladeck, David C. "Machines without principals: liability rules and artificial intelligence." *Wash. L. Rev.* 89 (2014): 117.
- Vollmann, Jochen, and Rolf Winau. "Informed consent in human experimentation before the Nuremberg code." *BMJ: British Medical Journal*313, no. 7070 (1996): 1445.
- Walter, Jennifer K., and Eran P. Klein. *The story of bioethics: from seminal works to contemporary explorations*. Georgetown University Press, 2003.
- Weiner, Charles. "The recombinant DNA controversy: Archival and oral history resources." *Science, Technology & Human Values* 4, no. 1 (1979): 17-19.
- White, Ronald F. "Institutional review board mission creep: The common rule, social science, and the nanny state." *The Independent Review* 11, no. 4 (2007): 547-564.
- Winham, Gilbert R. "International regime conflict in trade and environment: the Biosafety Protocol and the WTO." *World Trade Review* 2, no. 02 (2003): 131-155.
- Wolfensohn, Sarah, and Maggie Lloyd. *Handbook of laboratory animal management and welfare*. John Wiley & Sons, 2008.
- World Health Organization. *Laboratory biosafety manual*. World Health Organization, 2004.
- World Medical Association. "World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects." *JAMA* 310, no. 20 (2013): 2191.
- Yampolskiy, Roman, and Joshua Fox. "Safety engineering for artificial general intelligence." *Topoi* 32, no. 2 (2013): 217-226.
- Yudkowsky, Eliezer. "Artificial intelligence as a positive and negative factor in global risk." *Global catastrophic risks* 1 (2008): 303.

About FPF: Future of Privacy Forum is a nonprofit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. FPF brings together industry, academics, consumer advocates, and other thought leaders to explore the challenges posed by technological innovation and develop privacy protections, ethical norms, and workable business practices.

1400 Eye Street NW, Suite 450 | Washington, DC 20005 | FPF.org