http://social-epistemology.com
ISSN: 2471-9560

The Method of Convergent Realism

Chris Santos-Lang, Citizen Scientist, langchri@gmail.com

Santos-Lang, Chris. 2022. "The Method of Convergent Realism." *Social Epistemology Review and Reply Collective* 11 (1): 33-49. https://wp.me/p1Bfg0-6t5.

SERRC
Social Epistemology
Review & Reply Collective
11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

This essay seeks to advance a discussion to meet the needs of designers of technologies–including of institutions–which are meant to help users accurately answer questions about reality, including questions about nature and morality. Specifically, it helps clarify the list of requirements that designs would have to satisfy in order to provide reasonable expectation that the technology would converge on truth. The design of the procedures of the Belleville Research Ethics Committee (BREC) are offered as a practical example of how such a list of requirements might inform design, and this essay aims to help other designers reason about potential improvement to the BREC procedures.

Why Care and What to Hope For

In 2011, IBM's Watson artificial intelligence won Jeopardy because it could read much more than any human could. Now doctors around the world have "Ask Watson" buttons they can press to find out what treatment Watson recommends for their patient. IBM expects doctors to become dependent on such buttons as society's collective knowledge about health grows to exceed what individual doctors can read. Watson already follows over 300 medical journals and can read in over 13 languages. What human doctor can compete with that? But why should Watson stop with doctors? Why shouldn't I have an "Ask Watson" button to inform my decisions when planning my meals and exercise regimen? For example, when I eat at a restaurant, perhaps Watson could recommend items from the menu, factoring-in my personal health profile and eating history. Why shouldn't legislators have an "Ask Watson" button they can press to inform their votes on health legislation? Why shouldn't the button extend to legislation that impacts the health of our environment and economy? If we think doctors will need computers to inform their decision-making, why shouldn't we expect policy-makers to need the same? Why shouldn't I have an "Ask Watson" button that tells me which policy-makers to vote for in elections?

Even if Watson could help us be better voters, you might object that AI-guided voting would be pointless. If Watson, Siri, Alexa, Cortana, and Google all gave the same advice, and most voters trusted that advice, then outcomes would be decided before the vote, so it would be efficient to skip the actual voting ritual. Such profoundly informed voting might still be called "democratic," but it would not be the form of government Winston Churchill referred to when he said the best argument *against* democracy "... is a five-minute conversation with the average voter." One way such democracy would be new would be to manifest new vulnerabilities.

Much as one might anticipate threats to a nation's elections, this article anticipates threats to the social epistemic method that would guide decision-making in a technologically informed world. John Dewey referred to that epistemic method as "the scientific method" (1910), but it will instead be called the "method of convergent realism" herein because it can inform all kinds of decision-making, including moral and mathematical decision-making in so far as moral and mathematical facts are real.

All theories of convergent realism share two commonalities (e.g. Putnam 1982; Hardin and Rosenberg 1982):

- Supposition that natural laws/facts exist, and;
- Supposition that at least one practical method exists which would reliably converge on those laws/facts.

Rather than defend these suppositions, this article simply assumes them and focuses on the puzzle, "What would have to be true of the most-efficient reliable method of convergent realism?"

In the past, some philosophers proposed that "convergence" should be understood as requiring a method that yields a sequence of theories, each one more accurate than the previous (e.g. Popper 1963; Post 1971). Philosophers employing that definition were trounced in 1981 (Laudan). In contrast, the definition of "convergence" employed in the current essay requires merely that the method of convergent realism settles on truth in the end. We prefer whichever method gets there the fastest, but any method which settles on truth qualifies, even if it takes detours along the way.

To exemplify convergence, consider four proposed methods to find the exit of a finite labyrinth:

1. Wait for the answer to come to you;
2. Wander randomly;
3. Wander randomly, but at each fork prioritize whichever path you've travelled least;
4. Wander randomly, but build a map as you explore and prioritize paths you haven't explored yet.

The first method might not be reliable because one might wait forever. The other three methods involve randomness, so any of them could accidentally head away from the exit at some point (even if initially on the optimal path), and therefore fail to "converge" by the 1971 definition. By our new definition, however, the third and fourth methods both reliably converge on the exit, although the fourth is preferred over the third because it sometimes converges faster (and never slower).

Other proposals might need to be compared by applying them to sets of randomly generated labyrinths, but the four proposals listed above can be compared with mere reasoning: We can rule out the first two methods by recognizing the possibility of waiting forever and the (unlikely but real) possibility that random wandering may produce an infinite circle. We can recognize the inferiority of the third method compared to the fourth by recognizing a possibility like the following:

> You come to a fork. Turning right brings you to a second fork in ten feet. Turning right at the second fork brings you to a third fork in 500 miles.

SERRC
Social Epistemology
Review & Reply Collective
11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

> Turning right at the third fork brings you to the left branch of the first fork in ten feet. Reaching the second fork a second time, the method requires you to go left because that is the only path at that fork you have not already traveled at least once. It dead-ends in ten feet, so you must return to the second fork for a third time. Now the method requires you to repeat the 500 mile path to the third fork because that is the only path you have not travelled at least twice. Clearly, this is inefficient—you already know that path goes to the third fork, and you know that you could instead get there in twenty feet by going the other way.

Like mathematicians who take years to find a proof but recognize it in minutes once it is brought to their attention, you might not see the inefficiency of the third method until a possibility like the one above is brought to your attention. The rest of this article brings similar possibilities to the reader's attention--ways one could (inadvertently) design "Ask Watson" buttons that could lead society astray. Each potential for disaster implies a restriction on what the ideal method of convergent realism would have to entail in order to mitigate that potential.

This article does not assume that the ideal method of convergent realism aligns with current practice in science. On the contrary, this essay is sympathetic with modern efforts to reform science. Furthermore, this article does not assume with Dewey that every step of the method must be implemented by humans. Today, we must confront the possibility that machines might implement some steps better than humans. This essay also entertains the possibility that certain steps might best be reserved for certain *kinds* of humans--a division of labor which can make the ideal method necessarily social (Kitcher 1990).

By ruling-out alternative variations of the method of convergent realism, this article concludes that the ideal method of convergent realism must include at least these three steps:

1. Articulation of reasoning;
2. Independent tests of the reasoning;
3. Provision of means by which new discoveries will force retesting:
    a. Transparency;
    b. Conflict resolution;
    c. Amendment process;
    d. Expiration dates;
    e. Objectivity.

Any flaws in this conclusion could be expressed in terms of counter-examples in which ethics, mathematics or science can reasonably expect to converge on truth despite skipping one or more of these steps. Anyone who is aware of any such counterexample is asked to please share it via the PubPeer page for this article so that all readers can find it. Any reader of this article is asked to please check that page before trusting anything herein. Thus, the value of this article may be less to directly guide a redesign of modern AI and human

pursuits of truth than to open a productive conversation that can guide such design going forward.

In addition to justifying each step of the method in terms of the potential disasters it would avert, this article will provide examples to demonstrate the feasibility of each step. In particular, the author has constructed and implemented a full set of governance documents for the Belleville Research Ethics Committee (BREC). They serve as a "proof of concept" in the domain of research ethics.

Step 1: Articulation of Reasoning

Reasoning has two parts: (a) a claim and (b) what Longino (1990) called "background assumptions." In science, the claim is composed of a testable hypothesis, and the background assumptions justify methods used to test that claim. For example, reasoning might be composed of a claim about nitrogen plus background assumptions that include the assumption that a certain method is appropriate to purify the nitrogen used to test that claim. The discovery of a more reliable or precise way to purify nitrogen could challenge the background assumption part of that reasoning which could, in turn, change what we believe about the claim by obliging us to repeat its test with more reliably pure nitrogen. In mathematics, the claim is composed of a conclusion, and the background assumptions are composed of axioms and lemmas cited to justify each step in mathematical arguments. For example, the axioms of Euclidean geometry are background assumptions for the Pythagorean Theorem.

In ethics, the claim is a prescription, and the background assumptions include a supposed complete list of relevant ethical considerations such as:

- Does the prescription treat others with respect?;
- Is the prescription consistent with other accepted prescriptions (e.g. laws, traditions, religious beliefs)?;
- Compared to alternative prescriptions (or doing nothing), what consequences would following the prescription have for the well-being of humans alive today?;
- What consequences would following the prescription have for the well-being of biological, social or economic ecosystems?

The moral quality of a prescription is tested by application of the list of considerations. If a list were not complete–for example, if it excluded consideration of impacts on entities with whom we cannot empathize so directly (e.g. on the economy, on our ecosystem, on future generations, etc.)–then an initial endorsement of the prescription might be challenged by pointing-out the relevance of excluded considerations. The possibility that lists may be incomplete makes moral knowledge subject to revision (but with potential to converge) just like scientific knowledge.

11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

In *research* ethics, the claim would be a prescription that a certain research plan should be followed; the background assumptions would be a supposed complete list of relevant considerations. For example, before conducting the first research on radioactive chemicals, the researchers might have considered whether a different research plan would answer the same questions more efficiently. The subsequent discovery that radiation can cause harm produced a new background assumption for research ethics: consideration of whether the research plan includes sufficient protocols to mitigate risks of harm due to radiation. From then on, scientists became obliged to additionally consider radiation risks, which made them augment their experiments with new safety protocols (or, in some cases, stop doing a given experiment at all).

The method of convergent realism does not require that new claims contain previous claims, but it does require that new reasoning addresses previous background assumptions. For example, once significantly better procedures to purify nitrogen become discovered, any scientist who purified nitrogen in the obsolete way would be deviating from the method of convergent realism (unless they could justify the use of the obsolete method for the given case). The same would be true of any scientist who failed to account for radiation risks once radiation's potential to harm was discovered. It is by expanding and refining background assumptions that science, ethics, and so forth converge on truth.

To show that the ideal method of convergent realism must include articulation of reasoning, imagine an example in which Watson is fed claims without articulated background assumptions. Imagine a researcher is planning research and asks Watson whether implementing their plan would be ethical. Imagine Watson has read the published opinions of three distinct ethics committees who have already considered an equivalent research plan; one rejected the proposal for failing to mitigate radiation risks, but the other two approved the plan without considering radiation risks. If the published opinions did not articulate what ethical considerations each committee made, then Watson could only report, "Two out of three committees consider this plan ethical" when it should instead say "The morality of the plan may depend upon the relevance of radiation risks. Only one committee considered radiation risks, and that committee rejected this plan."

As another example, imagine someone asked Watson to report the current best estimate of a measurable property of nitrogen and Watson had read the reports of three teams who measured that property; none of the researchers knew how their nitrogen was purified (each simply acquired a cylinder of nitrogen from some supply center), and the team whose nitrogen was purified in a more precise way got a lower result than the other two teams did. If the scientists did not report how the nitrogen they measured was purified, then Watson would report a weighted average of all three teams' results, when it may be more appropriate to report only the results of the team that used the most pure nitrogen.

Ideally, the first time a certain kind of background assumption appears in reasoning, such as the first time anyone reported use of a more reliable way to purify nitrogen, Watson should automatically ask previous reporters of the same or contrary claims whether they dispute the relevance of the new background assumption. If the previous claimants do not dispute the

relevance, it should ask them to add the same background assumption to their own report (i.e. repeat their measurement with the better purification process). It should also tell teams–before they execute a research plan–what background assumptions will be expected in their report (i.e. "if you plan to report on that claim, be sure to purify your nitrogen *this* way…")

This essay demonstrates the feasibility of Step1 using the procedures of the Belleville Research Ethics Committee (BREC) as the example. The BREC procedures require investigators to articulate their ethical reasoning in three sections:

> 1. **Facts** include the research plan and any other information future investigators or committees might use to determine similarity to future cases;
> 2. **Precedents** list previously published opinions for similar cases;
> 3. **Considerations and Reasoning** list all ethical considerations made, explain how each was addressed, and defend any deviations from precedent.

Note that the BREC procedures place the responsibility to articulate ethical reasoning on those who propose research, rather than on the ethics committees who review it. In the future, it is doubtful that the method of convergent realism must necessarily delegate all of this responsibility to the individuals who make a claim. For example, perhaps individuals could ask Watson to find precedent, and to list all background assumptions articulated in precedent as a sort of checklist of what to consider with respect to their own research plan.

The first opinion authored by BREC cited 26 ethical considerations and 2 precedents (Committee Opinion about Replication of Merolla et al. 2017). The research plan was a replication study, so the first precedent was the opinion of the board which approved the original research. Instead of listing the ethical considerations it made, that board offered a letter noting that all approved plans must comply with "the Belmont principles, 45 CFR 46, and pertinent OHRP guidance" (Gerstein 2016). At least half of the ethical considerations listed in the BREC opinion went beyond that list of policies (partly because those policies intentionally ignore all ethical considerations beyond a scope called "human subjects"). Having better background assumptions made the BREC review more thorough than previous reviews (i.e. converging towards truth).

Furthermore, many of the 26 ethical considerations included in the first BREC opinion are relevant to many other research projects. Each research plan is different, so the list of considerations made by the next researcher might not include all 26 in BREC's first review and also might include considerations beyond that list, but we should not be surprised if 99% of the millions of research plans implemented each year would be covered by the 50 most common ethical considerations. If so, convergence would have a sense of "settling" as the average number of new ethical considerations invented per research plan approaches zero and stays there. By incorporating information technology, an economy of scale can be achieved which makes articulation of reasoning very feasible.

It is worth noting that the BREC procedures described above are *innovations* inspired by formalization of the method of convergent realism. As far as we know, previous research

11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

ethics committees recorded merely "approved" or "disapproved" for each plan they reviewed. (Some probably recorded more, but will not disclose those records–this essay defers further discussion of transparency until Step 3). As explained above, such records could mislead Watson because they do not specify which ethical considerations were made (or failed to be made). Citing the Belmont principles, 45 CFR 46, pertinent OHRP guidance, or more-complete lists of consideration to be made as records of how ethicists *actually* reasoned assumes that those ethicists never made any mistakes (which is not plausible). Without amassing reliable precedent, even manual research ethics must constantly reinvent the wheel, and that makes it unreasonable to expect research ethics to converge on truth without reform. In this sense, the BREC procedures are an example not only that Step 1 is practically feasible, but also that it is not trivial–there is a real need to raise awareness of it.

Step 2: Independent Tests of the Reasoning

In science, the testing step consists in observing *independent* examples in which the claim manifests or does not. Often this takes the form of a controlled experiment and replications of it, but astronomy is a field in which hypotheses are created and tested empirically even though controlled experiments are impractical. In mathematics, the testing step consists in submitting the reasoning to *independent* peer-reviewers. In ethics, it consists in submitting the reasoning to some kind of *independent* committee or court (which may have a single judge, multiple judges, a jury, or even a court of public opinion).

Independence is relevant in all of these cases because all of these tests are fallible. Tests qualify as independent from each other to the extent that their differences allow them to correct for each other's biases. Inability to correct for such bias is the problem with a mathematician providing their own peer-review, with a scientist collecting non-independent samples, or with a single ethicist serving as prosecutor, judge, *and* jury.

Independence may never be complete, but the method of convergent realism needs only enough independence to converge on truth eventually. In ethics, if a court fails to discern moral truth, then its opinion is expected to be overturned by a higher court or by the court of an independent (potentially future) community. In mathematics, if peer-reviewers fail to recognize an error, then it is expected that future readers will catch it. If an experiment yields a misleading result (which is statistically likely to happen from time to time), the error is expected to be caught via attempts to replicate it.

This essay held up BREC as a demonstration that it is feasible to implement Step 1 in research ethics by gradually increasing articulation of background assumptions over time. Likewise, it holds up BREC as a demonstration that Step 2 can be implemented in research ethics, gradually increasing independence of testing over time.

In the United States, special kinds of committees have evolved to provide independent tests for research ethics: Institutional Review Boards (IRBs) offer independent opinion about how well a research plan addresses considerations related to human subjects, Institutional Animal Care and Use Committees (IACUCs) offer independent opinion about how well the plan

addresses considerations related to animal subjects, and Institutional Biosafety Committees (IBCs) offer independent opinion about how well the plan addresses considerations related to recombinant DNA. To increase independence in opinions, all IRBs, IACUCs and IBCs already strive to include non-academics—often called "community members"—who represent the perspective of the general community. Inclusion of community members is expected to mitigate two risks:

> 1. Professional researchers are biased not to reject too many proposals, lest they lose their income stream, and;
> 2. Professional researchers are less independent from each other because they share similar training.

Typical IRBs, IACUCs and IBCs currently have difficulty with including community members (Klitzman 2012), but BREC-style procedures address the obstacles to inclusion.

The first obstacle to including regular people seems to be that many regular people find it intimidating to be a minority in a crowd of professional scientists. Imagine being expected to cast the deciding vote in a meeting full of professors who passionately debate each other using technical vocabulary you don't understand. Traditional IRBs, IACUCs and IBCs recruit enough professional scientists to cover the full range of specializations relevant to all research proposals, and such a large number of professionals inevitably makes community members a minority. In contrast, BREC is composed mostly of community members because it accesses specialist expertise relevant to a given proposed plan via temporary consultants. Furthermore, the BREC procedures require all members to agree to a set of expectations which include the expectation that "Expert members of the Committee are expected to teach non-expert members whatever is needed to form their own legitimate opinions." Each BREC ethics review serves the dual purpose both to advance knowledge at the expert level and to disperse expertise into the non-expert community.

Liability can be a second obstacle. People have proven willing to volunteer in many capacities, but few will volunteer to be held responsible if anything goes wrong, and that makes people hesitate to volunteer for an ethics committee. The Health Care Quality Improvement Act of 1986 gave medical peer-reviewers qualified immunity from liability so they would be more willing to offer independent review. Hoffman and Berg (2005) recommended giving similar immunity to IRB members for the same reason. But, even if an IRB member can't be sued, they might still *feel* responsible for harm done by scientists "under their watch." The problem is that scientists delegate ethics to an IRB the way patients delegate health decisions to their doctors. In contrast, the BREC procedures require scientists to fully understand their own ethics. The BREC procedures make the review process entirely transparent to scientists, and require scientists to seek additional independent review until the *scientists* determine that review was conducted adequately.

Commitment-level can be a third obstacle. Most non-commercial IRBs, IACUCs and IBCs serve a specific laboratory, university or hospital, and review all research at that institution. The commitment to review all research an institution conducts requires the average

11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

community member to review one plan per day. This burden is too high for many people. In contrast, BREC expects to review only one plan per year, and the BREC procedures empower its Chair to keep that burden low by rejecting additional requests for opinion. Investigators who cannot be served by BREC are given step-by-step instructions to launch additional BREC-style committees (https://goo.gl/LZr5nS).

One potential criticism of BREC-style committees compared to the current status quo is the worry that BREC-style committees might not provide assurance that all research will get reviewed (since BREC-style committees can reject requests for opinion). One answer to this criticism is that society can have *both* kinds of committees, so BREC-style committees don't need to provide everything that other research ethics committees can provide. Another answer is that traditional research ethics committees might have even less potential to provide such assurance. Limited supply of expert labor sets a cap on the number of proposals traditional IRBs can review non-superficially, and that cap may already have been reached (US Department of Health and Human Services 1998). IRBs should be expanding to include more and more ethical considerations (e.g. beyond human subjects). Instead, policy-makers are advancing policies to exempt more research plans from review (US Department of Health and Human Services 2017).

One famous answer to the problem of limited supply of expert labor is citizen science. As an example, faced with a classification challenge that experts could not complete in their own lifetime, the Galaxy Zoo project recruited 100,000 volunteers who classified more than 40 million galaxies in 175 days (Lintott et al. 2010). It remains to be seen whether hundreds of thousands of volunteers would likewise rise to the challenge of providing independent ethics review for every research plan scientists propose, but BREC-style procedures provide a way they could. We would have no reason to expect research ethics to converge on moral truth without non-superficial independent review, and that may require a flood of committees like BREC.

Other areas of ethics, or mathematics, or science could face the same obstacles to independent testing that BREC overcomes. For example, as science becomes more prolific, it might become impractical to independently test the replicability of each experiment without volunteer help. The most efficient method of convergent realism may need to bridge the gap between expert and non-expert scientists and mathematicians, and the similar gap between humans and AI. The competence of experts and AI both rely on independent testing, and the paths to assure independent testing may require that they elevate more of their community to similar competence.

Step 3: Provision of Means by which New Discoveries will Force Retesting

Science, mathematics, and ethics each have a history in which some claims which were tested many lifetimes ago are now considered outdated. New background assumptions became available, and retesting revealed that the Earth is not the actual center of the universe, geometry is not necessarily Euclidean, and moral authority does not actually derive from gender or bloodline. If such retesting never happened, then science, mathematics, and ethics

would converge on dogma, rather than on truth. One attempt to quantify this risk in science surveyed articles published in the top 100 psychology journals and found that only about 1.6% of the articles documented retests, yet retests by the original authors failed 10% of the time (Makel, Plucker, and Hegarty 2012). The implied convergence on 90% truth (or worse) is called the "replication crisis" because retesting is the implied solution.

This supposed "replication crisis" highlights the fact that being subject to retesting is not the same as *actually* getting testested. While Step 2 focuses on the independence of retests, Step 3 focuses on making sure those retests actually happen. Critical retests may require new background assumptions which were not available in the lifetime of the original researchers, so how could researchers possibly be responsible for Step 3? It is through "responsibilities of citizenship" that researchers maintain the fields which will retest their reasoning as background assumptions change. Some of the acts of good citizenship these responsibilities demand of a researcher may seem unrelated to that researcher's specific claims, but it would be unreasonable to expect to converge on truth if the responsibilities of citizenship are not met, and therefore would be unreasonable to make any claim without performing acts of good citizenship.

Specifically, this section argues that the following five acts of good citizenship are necessary to accomplish Step 3: transparency, conflict resolution, amendment process, expiration dating, and objectivity.

Act 1: Transparency

Transparency is about how well reasoning is explained. If the author of a claim is the only person who can understand its reasoning, then independent retesting becomes impossible, and articulation of background assumptions becomes useless. Even less extreme degrees of opaqueness can undermine the reasonableness of expecting convergence. For example, even a well-written explanation would be practically useless if lost in a fire or buried in "the Internet." BREC-style procedures require publishing opinions and all related materials (with any sensitive information redacted) to the Open Science Framework (OSF), where they are encrypted and backed-up twice a day to multiple locations. The OSF is backed by a $250,000 preservation fund which is expected to ensure free access to its data on the World Wide Web for at least fifty years. All of these details are important. To be transparent includes making one's publication intelligible, accessible, discoverable, citable, and permanent (but revisable). Researchers may someday be able to achieve all of these aspects of transparency through direct communication with something like Watson, but, regardless of how transparency is achieved, it clearly would be unreasonable to expect Watson to converge on truths it never encounters.

BREC may have been the first IRB to publish its opinions so transparently. Here's why all research ethics committees should do the same: Suppose a researcher submitted a research plan for review, and the research ethics committee discovered an ethical problem with it, but kept that discovery between the researcher and the committee. While the committee prevented that researcher from causing harm, other researchers may stumble onto the same

11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

idea, find no record of your discovery, implement the unethical plan, and end-up inflicting the harm. The committee could have prevented more harm—maybe even saved lives—if it published its opinions more transparently.

One theme to notice in Step 3 is that acts of good citizenship often support research not "owned" by the actor. For example, transparency isn't just the act of the research ethicist who posts information to the Internet when they discover a reason not to conduct an experiment. Transparency is also the act of the truck drivers who deliver the materials to generate the electricity that powers the servers which archive and index that post a century after the research ethicist passed away. Transparency is becoming a group effort and a struggle that will outlast any of us individually.

Act 2: Conflict Resolution

The second act of good citizenship is ***conflict resolution***. As an example, consider the conflict Marc Edwards resolved between Lee Anne Walters and the city engineers of Flint, MI, over the quality of its water supply. Lee Anne was a resident of Flint, tested her water, and found it unsafe. She was not a professional scientist, but she asked others to conduct independent tests. While many other non-professionals replicated her results (i.e. Step 2 occurred many times), many professional scientists refused to get involved, and the Flint engineers refused to fix the water supply. The conflict was resolved when Marc Edwards, a professional scientist from Virginia, repeated Lee Anne's test and confirmed her conclusion. The Flint engineers then fixed the water supply to the satisfaction of both Lee Anne and Marc.

It is worth noting that the water supply was not fixed to the satisfaction of Scott Smith, another non-professional who tested the water, and that Scott and Marc therefore found themselves in a justifiably passionate dispute over whether Flint residents should be told not to bathe. Hygiene is important for public health, so whomever was wrong in this debate was endangering Flint residents. Marc became threatened with life-altering legal action, until Scott conceded that bathing would be safe. Further discussion about Scott will be deferred to the section about *objectivity*, but he is mentioned here to demonstrate that it would be unreasonable to expect the city engineers to trust every report from people who think they are conducting science and that it would be unreasonable to expect all professionals to subject themselves to the same danger Marc braved. Fortunately, many professional scientists are just as qualified to do what Marc did–the Flint situation had little to do with Marc's own research. But if none of them perform the act of conflict resolution, then Watson would be as stuck as the Flint engineers, unable to sort through the conflicting claims.

The BREC procedures do not demonstrate that it is possible to reliably generate peacemakers, but one of the ways in which the BREC procedures improve over traditional procedures is to provide a mechanism. In contrast, each traditional research ethics committee has a jurisdiction such as a school, hospital, or laboratory. If two such committees reach contrary conclusions about the same research plan, then the research will

be allowed in one jurisdiction but not in the other. Thus, a research plan with global risks (e.g. risk of creating a pandemic) could be forbidden in ninety-nine jurisdictions but permitted in the hundredth. Thus, the procedures of traditional research ethics committees converge on all research getting permitted somewhere eventually. In contrast, the BREC procedures state that BREC opinions can be suspended or terminated by "any credible committee," and provide simple instructions to form additional committees as needed to resolve conflicts. The concept of "credible committee" may bring the same frustration that the concept of "professional scientist" brought to Flint, but the BREC procedures at least allow a mechanism by which a risk of global harm could be blocked globally.

Act 3: Amendment Process

The third act of good citizenship, maintenance of an ***amendment process***, involves making it feasible to reverse one's claims. It can take effort to avoid reaching a point of no return. For example, environmental scientists maintain seed banks, isolate nature preserves, and protect endangered species so that any damage can be undone if accepted reasoning is later found invalid. The classic example in which ethics suffer for lack of an amendment process is when the government of a state, religion, or business cannot be sufficiently amended to avoid corruption, so that state, religion, or business finds itself at odds with truth.

A common dystopian vision for AI involves someone corrupting it, permanently bending the AI to its own agenda, thus preventing the AI from ever converging on certain truths. Typical strategies to avoid such exploitation boil down to decentralizing networks, as in blockchain or the World-Wide Web, such that coordination is achieved more through open standards than through any central database that a corrupting entity could control. BREC is architected this way. It has procedures for amending opinions and for amending procedures, but it takes a further step analogous to open source software: Its infrastructure can be "forked" with a single click.

In contrast, despite being digital, modern journals do not automatically amend articles when their cited sources are redacted. Furthermore, some industries do not quickly and appropriately amend themselves to match discoveries about climate change and ecosystem change. Some governments do not quickly and appropriately amend themselves to match discoveries about threats of contagion and the efficacy of masks. Some religions do not quickly and appropriately amend themselves to match discoveries about social polarization. Institutions that are not well-maintained will become corrupt, resist truth, and battle against anyone who does converge on the truths they resist, so no one can reasonably expect to converge on truth without maintaining (or limiting) institutions.

Act 4: Expiration Dates

The fourth act of good citizenship is to record the dates of tests and retests. Test dates imply ***expiration dates***: If the most recent retest occurred long ago, then one must ask whether better background assumptions have since become available. That is a good reason for researchers to date their publications, which is what most researchers do, but many

11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

researchers neglect to publish retest dates. For example, a given mathematical claim may be retested each year by hundreds of college-level mathematics classes (i.e. proving a theorem on a chalkboard), but its most recently *published* retest may be decades old.

If expiration dates were recorded, then AI like Watson could use those records to flag information as "expired" and to tell researchers (and funders) what needs to be retested. But, if expiration dates are not recorded, then Watson might forever hold a false claim that never gets retested.

No innovation is necessary to demonstrate the feasibility of tracking expiration dates. This is already standard in research ethics. Following these standards, the BREC procedures require all opinions to be renewed at least annually, and requires that the implied expiration dates be published on all consent forms. Such fast expiration may, in fact, be overkill for most research plans.

Act 5: Objectivity

The fifth act of good citizenship is ***objectivity***. Some people may think science and mathematics are objective by definition, but the relevant kind of objectivity is the range of people who can test a claim. For example, claims about whether a work of art invokes a particular feeling in a particular person are considered *subjective* because only that person can test those claims. Yet mathematical reasoning would likewise lack objectivity if only a rare supercomputer could test it, and science that relies on rare spaceships, supercolliders, or giant datasets may be interesting, but is relatively non-objective. We perform the fifth act of good citizenship by widening the range of people who can access equipment and facilities required to retest claims. It starts simply with making spaceships and other resources less expensive, but then progresses towards open-sourcing software and hardware and towards establishing community labs.

The section on conflict resolution deferred further discussion of Scott Smith until now. Scott initially disagreed with Marc Edwards about the quality of the water in Flint, Michigan. Marc was not allowed to independently retest Scott's claim because Scott's test of the water quality used new proprietary technology he had developed. That *sounds* fishy: It is possible that the new test doesn't work. However, it is also possible for someone to develop a better test, and the typical way to secure compensation for such an invention is to restrict others' access to it (i.e. by sacrificing objectivity). This is an example of how lack of objectivity can prevent an AI like Watson from converging on truth by blocking the AI from telling which of the two possibilities is the actual truth.

Objectivity isn't just about increasing access to new technologies. It's also about keeping costs low for technologies that are already accessible. For example, researchers themselves are a technology that seem plentiful today, but they could become rare without ongoing parenting, educating, growing of food, and maintaining of shelters, medical services, security and financing. If researchers became rare, then retesting would become expensive, and whatever errors exist in the data used by Watson and other AI might never get corrected.

Thus, perhaps the most common act of objectivity is to raise the next generation. The people who perform this act vastly outnumber the people who call themselves "researchers", but play an equally essential role in the method of convergent realism.

The BREC procedures create the equivalent of community labs for research ethics. Most IRBs rely on some volunteers, but the BREC procedures eliminate the non-volunteer labor by leveraging the Open Science Framework, so a remarkably wide range of people can afford to assemble an equivalent committee and retest any BREC reasoning. In contrast, the Association for the Accreditation of Human Research Protection Programs reported that the median IRB budget in 2016 was about $800,000 annually or $400 per experiment (2017). This practically limits retesting to wealthy or institutionally-supported researchers. The reasoning of centralized courts such as a federal agency, Congress, or Supreme Court is even less retestable, since it would be even more expensive to assemble an equally-qualified group of people. People who wish to converge on truth—rather than on something distorted by politics—thus have good reason to treat centralized courts as a last resort.

Units of Convergence

Calling transparency, conflict resolution, amendment process, expiration, and objectivity "acts of good citizenship" raises the question, "Citizenship in *what*?" The answer to this question might be labeled the "unit of convergence" because it is the entity that is to converge on truth. The method of convergent realism requires transparency and conflict resolution within that unit, but does not require transparency and conflict resolution beyond it. For example, if the unit were humanity, then the method would not require humans to be transparent to whatever aliens (if any) exist on the other side of the universe. Likewise, it would not require us to resolve conflicts with aliens, maintain alien institutions, or make retests affordable to aliens.

Analogously, a unit of convergence could be smaller than the whole of humanity. Step 1 merely requires that it include parts that can articulate reasoning. Meanwhile, Step 2 merely requires that its parts be sufficiently independent to mitigate each other's biases and Step 3 merely requires that each unit of convergence be able to retest all claims indefinitely. Thus, if separate nations or companies could coexist indefinitely, then it could be reasonable for each to expect to converge on truth independently. Likewise, if an AI or individual human could sustain itself indefinitely and contain enough diversity to mitigate all biases, then it might reasonably expect to converge on truth all by itself. However, it seems unlikely that the method of convergent realism would settle on such individualism.

The pressure against individualism holds an analogy to the dynamics hypothesized in integrated information theory. According to this theory, consciousness comes in differently sized grains, such that consciousness of warmth might be of smaller grain than consciousness of a book because it requires more mental resources to recognize something as being a book (Tononi et al. 2016). The evidence to date concurs that removing pieces from a person's brain does not eliminate their consciousness, but does reduce the number of things of which that person can be conscious. Much as different grains of consciousness

SERRC
Social Epistemology
Review & Reply Collective
11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

may require different amounts of mental resource, different claims are retested at different costs (i.e. some require a supercollider, larger sample size, etc.). Thus, *assuming maximum efficiency*, a larger unit of conversion can reasonably expect to converge on more truths simply because more retests fall within what it can afford.

The "maximum efficiency" caveat above is significant. Each citizen almost constantly impacts the transparency, conflict resolution, amendment process, expiration, or objectivity of their unit of convergence. A citizen might not participate in *every* step of the method of convergent realism, but there are steps like "raising the next generation" (a part of objectivity) where fields overlap, where we cannot avoid having impact, and where one's only choice is how positive or negative one's impact will be. Some potential citizens might not even have much choice about that–they might not be able to contribute positively. For example, humanity as a whole includes some individuals who are too developmentally handicapped to learn language. Even tools like Watson cannot empower such individuals to overcome their handicaps, so they become excluded from the unit of convergence, at least regarding some claims. They are treated like children for their entire lives, never receiving complete transparency, never having their conflicts taken seriously, and never really being empowered to test certain claims for themselves. Regarding the most "advanced" claims of science, mathematics, and ethics, it might be argued that is how experts treat even the average human being.

Some elitism is unavoidable; we would go bankrupt trying to empower every member of every species to retest every claim we currently hold as true. On the other hand, if soybeans nourish independent testers, that might qualify soybeans as citizens of our unit of convergence, even though their contributions to convergence would be highly specialized. Soybeans and humans are not "equal" citizens in the sense of having equal opportunity to contribute in each role, yet soybeans may contribute more than certain humans, even more than certain brilliant humans. That's because a unit can be forced to exclude a potential citizen simply because that potential citizen's *arrogance* blocks that potential citizen from being transparent with, resolving conflicts with, or making retests affordable to other potential citizens. For example, imagine a nation that withholds strategic scientific discoveries or resources from the rest of the world. Such *arrogance* may force a unit to choose between potential citizens, thus escalating a conflict between individuals (or between individual nations) into a conflict between individualism and communalism (or nationalism and globalism). Conflict over what roles may be afforded to AI could likewise escalate into conflict between individualism (or speciesism) and communalism. Communalism survives such conflicts by undermining objectivity, increasing the dependency of individuals on communities. Yet we cannot reasonably expect to converge on truth when converging on subjectivity, so arrogance seems incompatible with reliable convergence on truth.

Conclusion and Potential Responses

This essay articulates a perspective that might not be shared by all readers. Some readers may doubt the existence of natural laws or facts of science, mathematics, or ethics. Others may doubt the existence of any reliable method we could follow to discern those laws or facts.

Finally, some may claim that there is a way to accomplish such convergence without following all of the steps mentioned here. Please share any such alternative method on the PubPeer page assigned to this essay so all readers can find them.

Yet other readers may consider the ideas in this essay trivial. "Of course we would have no reliable reason to expect Watson to converge on truth if the company feeding it new information will not last," they say, "That's why the Watson engineers chose to work at IBM, a company expected to last a long time." Yet even IBM might not last forever, so there are arguments to be made that Watson should be engineered to be transferable to an open source community, and to help humanity address existential threats such as climate change and nuclear stockpiling. This essay provides formalization to make such arguments rigorous. Moreover, comparing BREC procedures to those of traditional research ethics committees demonstrates that the formalization in this essay can be applied to yield practical innovation.

Other readers may share the perspective articulated here and appreciate its practical steps. They may believe it has meaningfully elaborated beyond previous attempts to describe a method of convergent realism. Ideally, they will share this elaboration among a unit of convergence, and elaborate even further.

References

Association for the Accreditation of Human Research Protection Programs, Inc. 2018. "2017 Metrics on Human Research Protection Program Performance for Academic Institutions." Washington, DC. https://admin.aahrpp.org/Website%20Documents/2017%20Academics%20Metrics.pdf.

Belleville Research Ethics Committee Procedures. 2017. IRB 11228. https://doi.org/10.17605/OSF.IO/CUXHB.

Committee Opinion about Replication of Merolla et al. 2017. IRB 11228. http://doi.org/10.17605/OSF.IO/SP8CM.

Dewey, John. 1910. *How We Think*. Boston, MA: DC Heath.

Gerstein, Dean R. 2016. Correspondence. ClaremontLetter.pdf (Version: 1) https://osf.io/h95yv/.

Hardin, Clyde L. and Alexander Rosenberg. 1982. "In Defense of Convergent Realism." *Philosophy of Science* 49 (4): 604-615.

Hoffman, Sharona and Jessica Wilen Berg. 2005. "The Suitability of IRB Liability." *Case Legal Studies Research Paper No. 05-4*. http://doi.org/10.2139/ssrn.671004.

Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* 87 (1): 5-22.

Klitzman, Robert. 2012. "Institutional Review Board Community Members: Who Are They, What Do They Do, and Whom Do They Represent? "*Academic Medicine: Journal of the Association of American Medical Colleges* 87 (7): 975–981.

Lintott, Chris, Kevin Schawinski, Steven Bamford, Anze Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert Nichol, Jordan Raddick, Alex Szalay, Dan Andreescu, Phil Murray, Jan Vandenberg. 2010. "Galaxy Zoo 1: Data Release Of

11 (1): 33-49. 2022.
https://wp.me/p1Bfg0-6t5

Morphological Classifications for Nearly 900,000 Galaxies." *Monthly Notices of the Royal Astronomical Society* 410 (1): 166-178.

Laudan, Larry. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* 48: 19-48.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Tradition and Change*. Princeton: Princeton University Press.

Makel, Matthew C., Jonathan A. Plucker, and Boyd Hegarty. 2012. "Replications in Psychology Research: How Often do They Really Occur?" *Perspectives on Psychological Science* 7 (6): 537-542.

Popper, Karl. R. 1963. *Conjectures and Refutations*. London: Routledge.

Post, H. R. 1971. "Correspondence, Invariance and Heuristics: In Praise of Conservative Induction." *Studies in the History and Philosophy of Science* 2: 213-255.

Putnam, Hilary. 1982. "Three Kinds of Scientific Realism." *The Philosophical Quarterly (1950-)* 32 (128): 195-200.

Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. "Integrated information theory: from consciousness to its physical substrate." *Nature Reviews Neuroscience* 17 (7): 450-461.

US Department of Health and Human Services. 2017. "Final Rule Enhances Protections for Research Participants, Modernizes Oversight System." *HHS Press Office*. http://wayback.archive-it.org/3926/20170127095200/https://www.hhs.gov/about/news/2017/01/18/final-rule-enhances-protections-research-participants-modernizes-oversight-system.html.

US Department of Health and Human Services. 1998. "Institutional Review Boards: A Time for Reform." *Office of Inspector General Publication OEI-01-97-00193*. June G. Brown, Inspector General. Washington, DC. https://oig.hhs.gov/oei/reports/oei-01-97-00193.pdf.