



State of the art review

Ethical perspectives on data and software sharing in the sciences: A research agenda



Peter T. Darch*, Emily J.M. Knox

School of Information Sciences, University of Illinois at Urbana-Champaign, 501 East Daniel Street, Champaign, IL 61820, USA

A B S T R A C T

Data and software are critical components of scientific work. Increased data and software sharing promises many benefits for science. Many stakeholders are building infrastructure and implementing policies to promote sharing. However, sharing remains rare in practice. Attention must be paid to researchers' ethical perspectives on sharing to fully realize the promise of sharing and promote greater circulation of data and software and better uptake of infrastructure for data and software curation. This research presents an agenda for researching these perspectives, including characterizing and accounting for researchers' perspectives; examining how these perspectives shape decisions related to data and software sharing; and understanding how and why differences in perspectives arise and are contested, negotiated, and resolved in multidisciplinary scientific collaboration. This agenda will enable stakeholders to identify and resolve differences in ethical perspectives, and develop policies, infrastructures, and education that support existing ethical perspectives, and cultivate better ethical practices.

1. Introduction

Data and software are critical components of scientific work, and major investments are made in their production. Data and software produced for one purpose often have potential to be successfully reused for other purposes, thereby advancing scientific progress (Pasquetto, Randles, & Borgman, 2017). Access to data and software underlying scientific results can also allow detection of malpractice or fraud, thereby promoting trustworthy science. However, the full potential of investments in producing data and software is far from realized. These research products often remain inaccessible to, or unusable by, other researchers, and often get neglected, leading to irretrievable loss (Heidorn, 2008).

To facilitate improved data sharing and reuse, funding agencies and universities have devised policies, and made investments in implementing digital infrastructure, for data curation (National Science Foundation [NSF], 2010). To improve compliance with these policies, and design and uptake of infrastructure, many library and information science (LIS) scholars work to characterize challenges to sharing, and to identify possible solutions (Chao, Cragin, & Palmer, 2015; Howison & Bullard, 2016). Nevertheless, sharing is still not widely practiced in most scientific domains (Tenopir et al., 2015). Uptake of digital infrastructure remains poor, compliance with data management policies is frequently not enforced or at the bare minimum level required, and interpersonal exchange of data between researchers is rare (Wallis,

Rolando, & Borgman, 2013).

Researchers' ethical perspectives – that is, their perspectives on what constitutes correct behavior in a particular situation – play a significant role in guiding their behavior across all aspects of scientific practice (Merton, 1973). However, typically embedded in initiatives and studies about data and software sharing is an assumption that researchers make decisions about whether to share based on an analysis of perceived costs and benefits to themselves. To fully realize the potential of data and software sharing requires a shift away from viewing researchers as actors who make decisions about sharing in accordance with self-interest, and towards also addressing the role of their ethical perspectives on sharing.

This paper outlines a research agenda for studying these perspectives, in particular 1) characterizing and accounting for perspectives held by researchers; 2) exploring how researchers from different disciplinary and institutional backgrounds negotiate, contest, and resolve differences in perspectives in multidisciplinary collaborative work; 3) understanding how perspectives change over time, and why; and 4) understanding how these perspectives guide decisions made that affect subsequent possibilities for data and software sharing. This research agenda will guide key stakeholders to 1) identify and resolve differences in ethical perspectives between researchers; 2) cultivate these perspectives further; and 3) devise and implement policies and infrastructure that both support existing ethical perspectives, and cultivate better perspectives.

* Corresponding author.

E-mail address: ptdarch@illinois.edu (P.T. Darch).

2. Data and software sharing: promises and reality

Data, as defined by Borgman, are “representations of observation, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship” (2015, p. 28). Increasing circulation of data is widely regarded as desirable for science. Four rationales for data sharing are 1) improving the reproducibility of research, 2) making the results of publicly funded research available to the public, 3) enabling researchers to ask new questions of extant data, and 4) advancing the state of research and innovation through enabling new ways of doing science (Borgman, 2015).

These rationales, although developed with respect to data, also apply to scientific software. Researchers increasingly use software to process and analyze their data. In some cases, this software is proprietary. In other cases, researchers write their own code, or use (and modify) open source software. Rationales that promote data sharing can be adapted to argue that software used to produce scientific results should be available to other researchers, in the sense that researchers should avoid using proprietary software and instead make publicly available the code they write or modifications they make to open source software. Improved software accessibility promotes reproducibility (Atmanspacher & Maasen, 2016). It also leads to improved scientific outcomes for science by reducing the need for researchers to write code that has already been produced by other researchers, and enabling researchers who lack the necessary skills to produce their own code to find and use software produced by others (D. S. Katz et al., 2016).

Improving data and software sharing is an important objective across science, big data and little data domains alike. Practices in big data domains, such as astronomy and bioinformatics, involve the use of very large datasets, on the scale of megabytes or terabytes, and computationally intensive methods (Borgman, 2015). Although a big data domain, at first sight, appears to enjoy data abundance, it is still typically interested in promoting data and software sharing, especially to increase the scientific output of the domain, to justify the large amounts of money invested in the domain infrastructure, and to promote reproducibility and transparency in the domain. Little data domains, by contrast, are characterized by access to much smaller quantities of data. Many of these domains struggle to access data sufficient to pursue their major objectives, a situation that would be improved by promoting circulation of data and software within these domains (Darch & Borgman, 2016).

Data and software can be shared via many methods, including interpersonal exchange and uploading to personal websites or online repositories (Wallis et al., 2013). To facilitate improved sharing, funding agencies and universities have made significant investments in developing data and software repositories (Clark, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2015; Steinhart, 2014).

Key stakeholders are also devising policies to promote sharing. One example is the NSF requirement for data management plans (NSF, 2010), which includes software as part of the definition of data. Some journals also require authors to make supporting data and software openly available—either as supplementary material or in a publicly accessible repository—as a condition of publication (Stodden, Guo, & Ma, 2013).

Despite these initiatives, and a stated willingness on the part of many researchers to share data and software, sharing is still not widely practiced in most scientific domains (Fecher, Friesike, & Hebing, 2015). Tenopir et al. (2015) found that the mismatch between stated willingness and actual behavior was particularly stark in the case of junior researchers.

Compliance with data and software policies is uneven. Data management policies are often underspecified and unenforced, leaving researchers with discretion about how to interpret and apply these policies in practice. Furthermore, bare minimum compliance with policies does not guarantee that the data and software will necessarily be accessible to, or reusable by, others (Zimmerman, 2008). The uptake of

infrastructure for data and software curation and circulation is also often patchy, and mere deposition of data and software is often insufficient to ensure these research products are usable by others.

3. Current research on data and software sharing

To inform the activities of initiatives undertaken to promote data and software sharing, LIS scholars have characterized barriers to data and software sharing.

3.1. Decisions that affect possibilities for data and software sharing

Researchers take many decisions, both explicitly and implicitly, that affect the degree to which their data and software can be, and are, shared. After producing data and software, researchers make decisions relating to how they might comply with any policies that apply to their work, and whether to use available infrastructure to deposit their research products. When faced with requests from other researchers for data or software, the researcher must decide whether and how to respond. Sharing with others requires the researcher to carry out work, such as cleaning datasets, adding metadata or documentation, and answering follow-up queries from potential re-users, to ensure the data or software can be interpreted and reused by others (Wallis et al., 2013).

In addition to decisions taken once datasets and software have been produced, decisions taken by the researcher at earlier stages of the research lifecycle, including stages of experimental design and data collection, also determine the extent to which data and software can be successfully shared (Wallis, Borgman, Mayernik, & Pepe, 2008). Choices made about how data are produced, collected, processed, and stored have a cumulative effect. For data sharing to proceed successfully, the potential re-user of the data must be able both to comprehend the data, and to judge its trustworthiness and reliability (Faniel & Jacobsen, 2010). Potential re-users of software may require detailed documentation about how the software was produced to adapt the software to their own purposes (Ince, Hatton, & Graham-Cumming, 2012).

The issues discussed so far apply to both data and software sharing. However, there are also some differences with data sharing that further complicate the work of software sharing (Trainer, Chaihirunkarn, Kalyanasundaram, & Herbsleb, 2015). Software sharing is a continuous process, even long after the initial sharing of the software occurs. For example, users may encounter bugs, which require the software producer to fix them. Producers of software also often respond to users' requests for new or more advanced features, which they may address in new versions of the software. Another complication with open source software is that users are able, and often encouraged, to build on the existing code to extend or modify the software. These adaptations place an additional burden on the original software producer, who may choose to incorporate these adaptations into future releases of the software.

A researcher must therefore make many decisions throughout the research lifecycle that affect subsequent potential for sharing data and software. Making data and software shareable places a significant burden on the researcher who produced them, often prohibitively so (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011), while the benefits to the researcher of sharing may be uncertain or non-existent.

3.2. Complicating factors in multidisciplinary research

The challenges and burdens of facilitating data and software sharing can vary from domain to domain, and multiply in the context of multidisciplinary research. Science is not a unified endeavor; instead it comprises a dynamic collection of evolving fields, with many variations in practices between fields (Galison, 1996). Some key examples of variations include methods and tools for generating knowledge, forms

of argument and reasoning; organizational structures and hierarchies, and size and scale of collaborative endeavors. Taken together, these (and other) variations give rise to distinct *epistemic cultures* (Knorr-Cetina, 1999).

Epistemic cultures often vary markedly in terms of issues related to access to and use of data and software, including the motivations for promoting greater data and software sharing. The ways in which domains differ have implications both for data and software sharing practices within individual domains, and when domains meet in multidisciplinary collaboration. Differences are particularly stark between big and little data domains.

Big data domains are typically characterized by access to, and use of, sophisticated digital infrastructure for data production, such as sky surveys or the Human Genome Project (Borgman, 2015). Datasets produced in this fashion are often assumed to belong to the community as a whole, rather than to any particular researcher or group of people. Many researchers write their own code for processing and analyzing datasets. Standards for data and software, such as metadata standards, formats, and documentation, are typically widely adhered to across a single big data domain.

Meanwhile, in little data domains, data and software sharing is inhibited by a number of factors particular to those domains (Borgman, Wallis, & Enyedy, 2007). Data are typically scarce and hard-won, and collected by an individual researcher or a small team, leading to researchers feeling a greater sense of ownership over the data they collect. The types, formats, and approaches to management of datasets can vary from researcher to researcher and infrastructure for data curation and circulation is typically patchy, impeding transfer and interpretation of datasets. Further, data very often get lost at the end of a project, or when the researcher leaves the domain (e.g., when a doctoral student moves into an industry position after completion of their degree). Scope for sharing software is often low because few researchers in the domain may write their own code to process and analyze data, relying instead on proprietary software.

Contemporary scientific practice is often conducted in collaborations comprising researchers from multiple domains (typically from two to six), each with their own distinctive epistemic culture, coming together to study a single phenomenon or system (Cummings & Kiesler, 2005). Although the challenges to promoting sharing data and software between two domains are obvious where one is a big data domain and the other is a little data domain, significant challenges can also exist at the interface of two big data domains. Even if each of the domains in question has well-established standards and infrastructure to facilitate data and software circulation internally, standards and practices are still likely to vary between the domains.

3.3. Studies and initiatives to promote data sharing

Typically embedded in the various initiatives and studies aimed at promoting sharing is an assumption that researchers make decisions about whether to share data or software based on an analysis of perceived costs and benefits (Trainer et al., 2015). Central to these initiatives and studies, therefore, is a concern with shifting the balance of burdens and incentives facing researchers.

Studies that focus on reducing burdens address resources available to reduce barriers to sharing (Tenopir et al., 2015; Wallis et al., 2013). These resources include repositories or registries for deposition; standards, for example, for data production methods and metadata (Lavoie, 2008), that allow for interoperability and ease of transferring data from one context to another; and funding and time available to carry out the work necessary to facilitate sharing. Increased provision of these resources can make data and software sharing less burdensome to realize.

Other studies and initiatives focus on changing incentive structures for researchers. Some studies address the issue of scholarly credit, for example, considering whether researchers will be more likely perform the work necessary to facilitate sharing if they receive some form of

scholarly attribution, such as citation (CODATA-ICSTI Task Group on Data Citation Standards Practices, 2013; Crosas, Carpenter, Shotton, & Borgman, 2013). More recently, attention has turned to devising principles for citing software (Smith, Katz, & Niemeyer, 2016). Credit for software is a particularly challenging issue because many individuals may have made incremental or slight contributions to a particular piece of software. Other studies highlight the effects of policies that require researchers to share data. These policies provide negative incentives, because sharing is required to avoid the risk of penalization (Fecher et al., 2015).

Despite these studies and initiatives and the uptake of infrastructure for and instances of sharing data and software, overall, data and software sharing still remains patchy. Note that the initiatives and studies discussed here largely focus on addressing researchers' self-interest. It is imperative that other determinants of behavior be explored.

4. Ethical decision making and scientific sharing practices

Viewing researchers as actors whose decisions are shaped largely by a calculus of perceived costs and benefits neglects considering the influence of their ethical perspectives. Rachels and Rachels define ethics as “the effort to guide one's conduct by reason—that is, to do what there are the best reasons for doing—while giving equal weight to the interests of each individual affected by one's decisions” (2015, p. 13). In other words, an individual's behavior is usually guided by considerations of other people in addition to him or herself.

Issues of ethics have long been known to play a critical role in shaping researchers' behavior (Kalleberg, 2007; Merton, 1973). Researchers are often concerned by the social impacts of their work, equitable distribution of credit, and research integrity. Ethical perspectives can thus be assumed also to play a significant role in shaping researchers' decision-making with respect to actions that affect data and software sharing.

Addressing ethical perspectives on sharing will help understand both researchers' willingness to undertake the work necessary to facilitate sharing, and their existing patterns of compliance and use of policies and infrastructure. For example, studies of relationships between ethics and technologies suggest technologies that require users to behave in ways contrary to their own ethical beliefs typically alienate potential users (van den Hoven, Vermaas, & van de Poel, 2015). These studies suggest that poor uptake of data and software infrastructure by researchers could be partly caused by conflicts with the researchers' own ethical perspectives about sharing. Understanding these perspectives can therefore inform the design of data and software infrastructure that are more attractive to potential users.

Hare (1981) argues that individuals' ethical reasoning can occur on two planes: the *immediate plane* and the *critical evaluative plane*. The immediate plane refers to judgments and actions based on an individual's ordinary moral sense, or their intuitive sense of what is right or wrong. The critical evaluative plane refers to judgments and actions based on deliberative reasoning about how to behave in a particular situation.

Kitchener and Kitchener (2009) expand the critical evaluative plane into four elements, producing a five-level model for ethical decision-making (with levels 2–5 corresponding to the critical evaluative plane):

1. Immediate level of moral reasoning;
2. Ethical rules, relating to rules such as those contained in codes of conduct;
3. Ethical principles, which are more general than rules or codes of conduct;
4. Ethical theory, providing a philosophical grounding for devising or justifying principles;
5. Meta-ethics, dealing with overarching issues such as how individuals reason ethically.

These five levels are intertwined. A component of a lower level can be justified by components of a higher level (e.g., ethical principles can provide justification for ethical rules, and also find a grounding in ethical theory). If one level of reasoning proves inconclusive, a higher level can be invoked to try and provide resolution (e.g., if different principles provide conflicting guidance about how to proceed, then an individual may turn to ethical theory to resolve this conflict).

The first four levels are discussed below: The fifth level (meta-ethics) cuts across the other four levels.

4.1. Immediate level of moral reasoning

An individual's moral sense develops as that individual takes cues from authority figures and other members of society about what is ethical. This sense further develops with experience of ethical reasoning about particular situations. This sense usually serves an individual well in familiar situations. However, in new situations an individual's moral sense may not provide an adequate guide.

The moral sense of a researcher develops in many ways. As their career unfolds, they learn about what is considered ethical in science from their mentors and advisors (e.g., doctoral advisors) and academic peers. However, data and software sharing is an emergent concern, and responding to requests to share may be an unfamiliar situation for researchers. Researchers may be used to collecting and managing data only for their own personal use, and not for possible reuse by others. Further complicating matters is the fact that these decisions about sharing data may need to be made at the interface of different scientific domains, where moral senses may conflict. In many instances relating to data and software sharing, then, a researcher's moral sense may not be adequate for guiding decision-making.

4.2. Ethical rules

Many professions have codes of conduct that set down a series of rules and ideals to which members should adhere. However, rules cannot completely determine behavior (Carusi & De Grandis, 2012). Reality is complicated and messy, and rules cannot take into account all the contingencies that exist in a particular situation. Individuals subject to these codes, or those who are charged with enforcing these codes, may still have significant discretion in terms of whether and how a code should apply in a particular situation.

Scientific researchers may encounter, or be subject to, codes of conduct or guidelines at different levels, for example:

- Individual laboratory or project;
- Institutional;
- Discipline-specific;
- General scientific associations, for example, the National Academy of Sciences' (NAS) "On Being A Scientist: A Guide to Responsible Conduct in Research" (National Academy of Sciences, 2009).

Topics of these codes include scientific integrity, human subjects research ethics, laboratory safety, public engagement, non-harassment and non-discrimination, and mentorship of early career researchers (Chadwick, 2005).

Occasionally, these codes cover issues related to data or software. For instance, some codes related to biomedical research direct researchers about handling data in a way that protects research subjects' privacy. However, many codes that address data or software are ambiguous about how researchers should behave in practice. For instance, while the NAS advises researchers to share data with other researchers upon request, it also acknowledges that data may be "too voluminous, unwieldy, or costly to share quickly" (NAS, 2009, p. 11). Further complicating the application of codes of conduct to decisions related to sharing is the fact that different researchers in a collaboration may be subject to different, even conflicting, codes (for instance, researchers

located in different universities, or in different disciplines).

4.3. Ethical principles

A number of ethical principles are relevant to scientific practice. Ambiguity and conflict can arise when these principles are applied to data and software sharing.

4.3.1. Mertonian norms of scientific practice: communalism and disinterestedness

The canonical examples of principles that apply to scientific practice are the norms set out by Merton (1973). The NAS (2009) cites these norms as inspiration. Merton argues that four norms characterize scientific practice, of which two are

- Communalism: Research belongs not to the individual but to the scientific community; and
- Disinterestedness: Scientists do not pursue their self-interest at the expense of the interests of the scientific community.

These two norms can apply to data and software sharing. Communalism suggests that data does not belong to an individual researcher, or group of researchers, but should instead be accessible to all researchers. Disinterestedness suggests researchers should not deliberately hoard data or refuse to release software to gain a competitive advantage over other researchers, if those data or software could be used by others to advance scientific research. However, subsequent research casts doubt on the claim that these norms underpin the behavior of scientists, suggesting researchers are indeed motivated by self-interest (Latour & Woolgar, 1979).

4.3.2. Beneficence and non-maleficence

Two principles related to each other are *beneficence* and *non-maleficence* (Beauchamp & Childress, 2001). Beneficence means that research activities should be conducted for the good of others, while non-maleficence means researchers should avoid activities that cause harm to others. These principles have concerned many researchers and ethicists. For instance, the advent of the atomic bomb provoked soul-searching on the part of many physicists, and motivated initiatives to apply physics to the medical sciences (Ceccarelli, 2001).

Beneficence can justify rationales advanced for sharing research data and software, because sharing can advance research and benefit society (Borgman, 2015). Meanwhile, non-maleficence can justify the reasons advanced by some researchers not to share data, in two ways in particular. The first is the concern that the shared data might be used for purposes the researcher considers nefarious (Wallis et al., 2013). For instance, climate science researchers may be concerned that their data will be misinterpreted or misrepresented by those with a vested interest in undermining climate science. The second, particular to human subjects research, is the risk of privacy breaches arising from data sharing. For instance, an individual's medical data, if made public, could be used by insurance companies or potential employees to discriminate against them (Narayanan, Huey, & Felten, 2016).

4.3.3. Justice

Another relevant principle is *justice*, namely the idea that benefits and burdens or harms of research should be fairly distributed across different social groups, such as socioeconomic class, race, gender, and nationality. This principle has arisen in many ethical issues related to science. One notorious example is the Tuskegee Syphilis Study, where researchers were found to have exploited a poor African-American community over 40 years, causing lasting harm to the health of many within the community (Katz, Capron, & Glass, 1972). Notions of justice also underpin concerns with ensuring that the benefits of technologies resulting from scientific research are distributed as widely or equitably as possible (Vayena & Tasioulas, 2016).

Justice can underpin arguments for data and software sharing. The open science movement, encompassing open data and open source software (alongside open access to scientific journals), is in part motivated by justice. The resources necessary to produce and use large-scale datasets are often prohibitive for researchers in underserved institutions, and researchers working in developing countries. One key argument of open science is that data and software sharing would allow these researchers to access resources for their research (Uhlir & Schröder, 2007). Justice also underpins the rationale, advanced by Borgman, that the products of publicly funded research, including data, should be made accessible to those who funded the research (i.e., taxpayers) (Borgman, 2015). This rationale balances the burden of paying for research with the right to access research products.

However, justice can also be used to justify not sharing data and software. For instance, justice grounds the idea that researchers who shoulder the burdens of producing data and software should be able to benefit from the data or software first, before having to share them with others (Wallis et al., 2013).

4.3.4. Honesty

Researchers should be honest when conducting research and presenting findings. They should not attempt to manipulate or falsify data or findings, or suppress inconvenient findings. Trust underpins scientific practice as the vast majority of research is not witnessed directly by others (Shapin, 1994). Fraudulent research has the potential to undermine both researchers' trust in each other, and the public's trust in science. Honesty is a key principle force driving the growing movement for reproducible science (Atmanspacher & Maasen, 2016), which entails data and software sharing.

4.3.5. Conflicting or uncertain principles

Invoking ethical principles may not fully resolve dilemmas related to sharing. Principles can provide conflicting guidance. For instance, disinterestedness suggests that researchers should serve science regardless of potential harmful outcomes, while pursuing non-maleficence requires researchers to act in a way that avoids harmful outcomes.

Different principles may apply depending on context. Conflicts of principle can multiply in multidisciplinary collaboration, as different principles may apply in different ways to different domains. Researchers in different domains may have different perspectives on data ownership, or on the goals of research in their domain. For instance, researchers in some domains may place a greater emphasis on the social impact of their research than those in other domains.

4.4. Ethical theories

Ethical theories, of which there are multiple types, ground principles.

4.4.1. Consequentialism and utilitarianism

Consequentialist perspectives involve an individual making judgments about a course of action on the basis of the anticipated consequences of that action. Utilitarianism, or doing that which causes the greatest happiness for the greatest number of people, is the most well-known consequentialist theory (Bentham, 1996).

Utilitarianism grounds principles of beneficence and non-maleficence. A justification for sharing based in utilitarianism would state that a researcher shares her data or software because the benefits of sharing (such as more efficient or new types of scientific research, or benefits to other researchers' careers if they can reuse data or software), outweigh the costs (such as potential for data to be misused, or burdens shouldered by the data producer in curating data to a standard where it can be easily shared and reused by others).

4.4.2. Deontologism and Kantian ethics

Deontological theories involve judgments about an action based on whether the act in itself is ethical, irrespective of its consequences. Kantianism is a form of deontologism. At its core, Kantianism is based on the categorical imperative, namely that one should treat others as ends in themselves, and not simply as means (Kant, 1996).

Kantianism grounds a number of principles. One example is honesty. For instance, deliberately falsifying scientific research can harm collaborators' reputations and waste other researchers' resources if they attempt to build upon the fraudulent research. From this perspective, data and software sharing can be seen as a way of demonstrating trustworthiness. Justice, too, has a Kantian grounding. For instance, researchers funded by public money should not simply see taxpayers as a resource to be exploited; instead, openly sharing data and software with the public can be regarded as a display of respect and reciprocity.

4.4.3. Contract theory

Contract theory is the idea that humans should hypothetically agree to norms of behavior necessary for the establishment and maintenance of social order. Hobbes (1998) argued that the natural human condition is a state of pure self-interest, leading to harmful outcomes for humans as they compete with each other for scarce resources. Instead, humans give up some individual rights to enable cooperation through the formation of larger social groups and institutions. By regarding science as an institution whose order relies on particular norms of behavior, Hobbesian theory grounds principles such as communalism and disinterestedness, which mandate that a researcher's conduct should privilege the wellbeing of science over their own personal interests. Thus, a researcher should not hoard data and software if sharing could help to advance science as a whole.

The philosopher John Rawls advanced a different version of contract theory. He argued that humans should derive ethical principles as though they do not know what position in society they occupy (Rawls, 2009). Rawlsian theory grounds principles of justice. For instance, a well-resourced researcher, when considering whether to share their own data, should reason as though they could equally likely be in a position where they do not have the resources to produce their own data. Hence, they should conclude that data sharing is the right thing to do.

4.4.4. Ethics of care

The ethics of care is a feminist critique of traditional ethics that holds that reason should not be prioritized over emotion when making ethical decisions and that there are particular decisions that are made based on relationships among persons (Held, 1998). The ethics of care places particular emphasis on nurturing wellbeing and development. This theoretical perspective could have a number of applications in science. Kinship networks are widespread in science, and include formalized relationships (such as doctoral students and their advisors) as well as informal mentoring networks (Murillo, Gu, Guillen, Holbrook, & Traweek, 2012). For instance, a faculty member who acts as an advisor to a student assumes a particular responsibility for the advisee's development as a researcher. In particular, the advisor may thus have a responsibility to share with the advisee (even if they choose not to share with others).

A second possible application of ethics of care is in the case of new or emerging domains of science. Very often, a single person, or small group of people, assumes a leadership role in establishing a new domain. Nurturing the domain is likely to involve promoting the production and circulation of data, to enable the domain to grow and produce scientific results, thereby establishing the credibility of the domain and strengthening it further. They are responsible for securing resources to enable the building of infrastructure and policies for data and software curation.

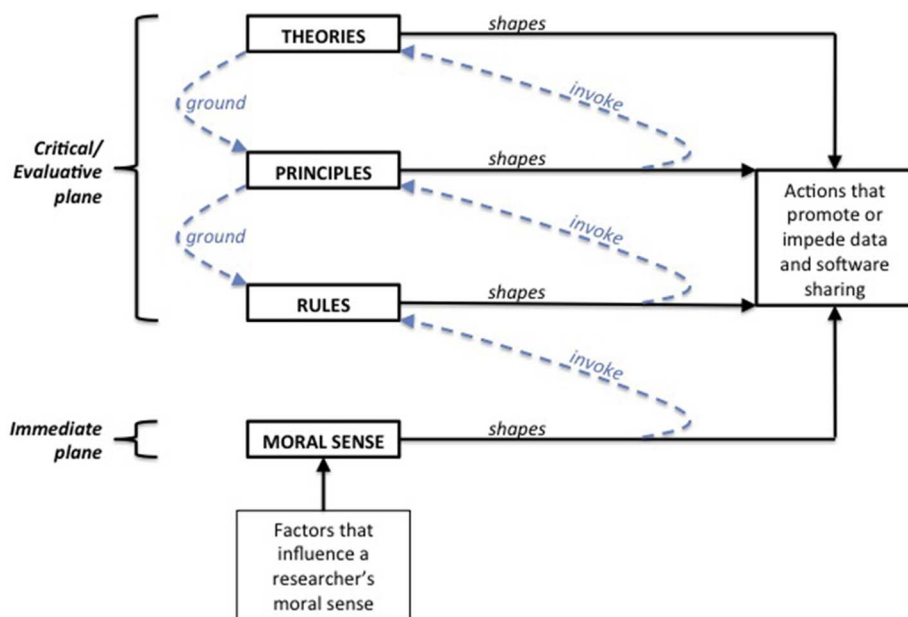


Fig. 1. A model for researching ethical perspectives on data and software sharing.

5. Research agenda for studying ethical perspectives on data and software sharing

The research agenda presented here is based on the model presented in Fig. 1. This model draws from the multi-level organization of ethical reasoning discussed in the previous section. It involves understanding researchers' moral senses, and how these moral senses shape decision-making processes in relation to actions (at all stages of the research lifecycle) that affect data and software sharing. It also involves understanding how and when moral sense is inadequate for shaping these decision-making processes, and thus where the critical evaluative plane becomes relevant. Each level in the critical evaluative plane can shape decision-making processes (Hare, 1981). Further, each of these levels provides grounding for lower levels, and may involve invoking higher levels. Issues of meta-ethics cut across the processes represented in Fig. 1 at all levels. Attention must be paid to separating and comparing perspectives on data sharing compared to software sharing, and to the purposes of sharing (e.g., sharing for reuse vs. sharing for reproducibility).

5.1. Immediate level of moral reasoning

This research agenda involves both characterizing and accounting for researchers' moral senses, and understanding how these senses shape data and software sharing practices.

5.1.1. Researchers' moral senses

Within a single laboratory or research collaboration, the immediate moral senses of researchers will often vary. It is important to study not only that part of a researcher's moral sense that relates directly to sharing practices, but to parts of their moral sense that relate to other aspects of scientific conduct: The researcher may not have had much opportunity to develop a sense in relation to sharing, but may instead draw upon these other parts of the moral sense.

Many factors shape an individual's moral sense, and its development over time. One factor is disciplinary background, given the disciplinary differences that exist regarding the role of data and software in research (Borgman, 2015). Characterizing a researcher's social and mentoring networks will also help account for their moral sense, for instance, by helping to understand how the sense of one researcher influences the sense of another, or to whom a researcher turns when they are trying to make ethical judgments around issues of sharing.

5.1.2. Moral senses shape decisions about sharing

Understanding how moral sense shapes behavior has a number of components. How an individual researcher applies their individual moral sense will affect sharing behavior. At the group level (laboratory, or collaborative project), it is important to study how the moral senses of group members interact, conflict, and are negotiated during decision-making processes, particularly in multidisciplinary settings. It is also important to understand where moral senses prove insufficient, and thus where researchers invoke, either explicitly or implicitly, the critical evaluative plane.

5.2. Critical/evaluative plane

For each level of the critical/evaluative plane, it is necessary to study what elements are invoked by researchers, how these aspects shape decision-making processes in instances related to data and software sharing, and what happens when these aspects are insufficient to completely resolve ethical dilemmas.

For instance, at the ethical rules level, it is necessary to study what codes of conduct or guidelines influence researchers. In some cases, these rules will explicitly relate to sharing; in other cases, their links to sharing will be tenuous and the researcher adapts them to the specific issue of sharing. In some cases, these rules will relate to the institution (e.g., discipline, university) in which the researcher is located; in other cases, the researcher will improvise rules devised for another institution.

The processes by which a researcher selects, adapts, rejects, and improvises ethical rules to guide their choices are worthy of study. Further, it is important to understand factors (e.g., disciplinary background, career stage) shaping these processes. The dynamics of how rules are compared, shared, contested, discarded, and accepted within group decision-making processes need to be addressed. Finally, attention should be paid to instances when rules prove insufficient for guiding conduct.

5.3. Infrastructures, policies, and training for data and software sharing

Using the research described above to improve data and software curation and sharing practices involves two approaches in particular, one relating to *descriptive ethics* (the process of describing and characterizing existing ethical perspectives held by researchers) and the other relating to *normative ethics* (which addresses questions about how

researchers should behave, and how to foster the desired behavior) (Kitchener & Kitchener, 2009).

5.3.1. Descriptive approach: interventions to accommodate existing ethical perspectives

A descriptive ethics perspective suggests considering how to devise, modify, and implement policies and infrastructure that better conform to researchers' existing perspectives (van den Hoven et al., 2015), including how to accommodate multiple perspectives simultaneously, not only increasing the range of researchers to whom the infrastructure and policies appeal, but also reducing the scope for conflict within groups of researchers. The patchiness of existing policies and infrastructure means there is much scope to intervene. These interventions could lead to increased uptake of infrastructure and better compliance with policies that go beyond the bare minimum necessary to satisfy requirements.

5.3.2. Normative approach: interventions to foster desired change in ethical perspectives

A normative ethics perspective suggests determining first, what ethical perspectives should be held by researchers, and second, how to foster these perspectives. The benefits of improved data and software sharing are widely agreed-upon. The challenge is thus to identify those elements of ethical thinking (senses, rules, principles, and theories) that promote activities that increase sharing, and then to devise interventions to lead researchers to develop these desired ways of thinking.

One component of fostering the desired change in researchers' perspectives is devising education and training, the needs of which may vary considerably across disciplines. A second component of fostering change is to devise policies and infrastructure that guide researchers to behave in ways that conform to desired ethical perspectives. Research on ethics and technology suggests that features and affordances of technology can shape people's ethical perspectives, by reinforcing certain types of behavior and discouraging others (van den Hoven et al., 2015).

6. Conclusion

Increased sharing of data and software is a major concern for a range of key stakeholders, including researchers, funding agencies, and governments (Borgman, 2015). Although ethics have long been held to shape researchers' behavior in general, the role of ethical perspectives on promoting or inhibiting data and software sharing has received little attention to date. The agenda presented here addresses these perspectives. It focuses on identifying differences in perspectives between researchers, why these differences arise, how perspectives change over time, and how differences are contested and resolved as individual researchers, laboratories, and scientific collaborations all make decisions that affect data and software sharing practices.

This research agenda promises to inform the work of scientific policymakers, scientific collaborations, and universities building infrastructure for data and software 1) to identify and resolve differences in ethical perspectives between stakeholders; 2) to cultivate ethical approaches to data and software management and sharing in scientists; and 3) to devise and implement policies and infrastructures that both support existing perspectives on ethics of data and software sharing, and cultivate better ethical practices. Understanding these ethical perspectives will enable investments in training and infrastructure to be targeted so as to be most effective, in turn ensuring that potential investments in data and software production made by public and private funding agencies is fully realized to the benefit of science and society.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Atmanspacher, H., & Maasen, S. (Eds.). (2016). *Reproducibility: Principles, problems, practices, and prospects*. Hoboken, NJ: John Wiley & Sons.
- Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford, UK: Oxford University Press.
- Bentham, J. (1996). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Oxford, England: Clarendon Press.
- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: The MIT Press.
- Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, 7(1–2), 17–30. <http://dx.doi.org/10.1007/s00799-007-0022-9>.
- Carusi, A., & De Grandis, G. (2012). The ethical work that regulations will not do. *Information, Communication & Society*, 15(1), 124–141.
- Ceccarelli, L. (2001). *Shaping science with rhetoric: The cases of Dobzhansky, Schrodinger, and Wilson*. Chicago, IL: University of Chicago Press.
- Chadwick, R. (2005). Professional ethics and the “good” of science. *Interdisciplinary Science Reviews*, 30(3), 247–256. <http://dx.doi.org/10.1179/030801805X42054>.
- Chao, T. C., Cragin, M. H., & Palmer, C. L. (2015). Data practices and curation vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology*, 66, 616–633. <http://dx.doi.org/10.1002/asi.23184>.
- Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. *Nucleic Acids Research*, 44(D1), D67–D72.
- CODATA-ICSTI Task Group on Data Citation Standards Practices (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, 12, CIDCR1–CIDCR75. <http://dx.doi.org/10.2481/dsj.OSOM13-043>.
- Crosas, M., Carpenter, T., Shotton, D., & Borgman, C. L. (2013). Amsterdam manifesto on data citation principles. Retrieved from <http://www.force11.org/AmsterdamManifesto>.
- Cummings, J., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, 35(5), 703–722. <http://dx.doi.org/10.1177/0306312705055535>.
- Darch, P. T., & Borgman, C. L. (2016). Ship space to database: Emerging infrastructures for studies of the deep subseafloor biosphere. *PeerJ Computer Science*, 2, e97.
- Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5), 667–690. <http://dx.doi.org/10.1177/0306312711413314>.
- Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Journal of Computer Supported Cooperative Work*, 19(3–4), 355–375. <http://dx.doi.org/10.1007/s10606-010-9117-8>.
- Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS One*, 10(2), e0118053. <http://dx.doi.org/10.1371/journal.pone.0118053>.
- Galison, P. (Ed.). (1996). *The disunity of science: Boundaries, contexts, and power*. Palo Alto, CA: Stanford University Press.
- Hare, R. (1981). The philosophical basis of psychiatric ethics. In S. Block, & P. Chodoff (Eds.). *Psychiatric ethics* (pp. 31–45). Oxford, UK: Oxford University Press.
- Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, 57, 280–299. <http://dx.doi.org/10.1353/lib.0.0036>.
- Held, V. (1998). Feminist transformations of moral theory. In M. Gatens (Ed.). *Feminist ethics* (pp. 3–26). Brookfield, VT: Ashgate/Dartmouth.
- Hobbes, T. (1998). *Leviathan*. Oxford, England: Oxford University Press.
- Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67, 2137–2155. <http://dx.doi.org/10.1002/asi.23538>.
- Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, 482(7386), 485–488.
- Kalleberg, R. (2007). A reconstruction of the ethos of science. *Journal of Classical Sociology*, 7(2), 137–160.
- Kant, I. (1996). *Kant: The metaphysics of morals*. Cambridge, England: Cambridge University Press.
- Katz, D. S., Choi, S.-C. T., Wilkins-Diehr, N., Hong, N. C., Venters, C. C., Howison, J., ... Littauer, R. (2016). Report on the second workshop on sustainable software for science: Practice and experiences (WSSPE2). *Journal of Open Research Software*, 4(1), <http://dx.doi.org/10.5334/jors.85>.
- Katz, J., Capron, A. M., & Glass, E. S. (1972). *Experimentation with human beings: The authority of the investigator, subject, professions, and state in the human experimentation process*. New York, NY: Russell Sage Foundation.
- Kitchener, K. S., & Kitchener, R. F. (2009). Social science research ethics: Historical and philosophical issues. In D. Mertens, & P. Ginsberg (Eds.). *The handbook of social research ethics* (pp. 5–22). Thousand Oaks, CA: SAGE Publications, Inc.
- Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge, MA: Harvard University Press.
- Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Beverly Hills, CA: SAGE.
- Lavoie, B. (2008). PREMIS with a fresh coat of paint: Highlights from the revision of the PREMIS data dictionary for preservation metadata. *D-Lib Magazine*, 14(5/6), <http://dx.doi.org/10.1045/may2008-lavoie>.
- Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, IL: University of Chicago Press.

- Murillo, L. F. R., Gu, D., Guillen, R., Holbrook, J., & Traweek, S. (2012). Partial perspectives in astronomy: Gender, ethnicity, nationality and meshworks in building images of the universe and social worlds. *Interdisciplinary Science Reviews*, 37(1), 36–50.
- Narayanan, A., Huey, J., & Felten, E. W. (2016). A precautionary approach to big data privacy. In S. Gutwirth, R. Leenes, & P. De Hert (Eds.). *Data protection on the move: Current developments in ICT and privacy/data protection* (pp. 357–385). Amsterdam, Netherlands: Springer.
- National Academy of Sciences (2009). *On being a scientist: A guide to responsible conduct in research*. Washington, DC: National Academies Press.
- National Science Foundation (2010). *NSF data management plans*. Washington, DC: National Science Foundation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp.
- Pasquetto, I., Randles, B., & Borgman, C. (2017). On the reuse of scientific data. *Data Science Journal*, 16(8), 1–9.
- Rachels, J., & Rachels, S. (2015). *The elements of moral philosophy*. New York, NY: McGraw-Hill.
- Rawls, J. (2009). *A theory of justice*. Cambridge, MA: Harvard University Press.
- Shapin, S. (1994). *A social history of truth: Civility and science in seventeenth-century England*. Chicago, IL: University of Chicago Press.
- Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, 2, e86.
- Steinhart, G. (2014). An institutional perspective on data curation services. In J. M. Ray (Ed.). *Research data management: Practical strategies for information professionals* (pp. 303–323). West Lafayette, IN: Purdue University Press.
- Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS One*, 8(6), e67111. <http://dx.doi.org/10.1371/journal.pone.0067111>.
- Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One*, 10(8), e0134826.
- Trainer, E. H., Chaihirunkarn, C., Kalyanasundaram, A., & Herbsleb, J. D. (2015). From personal tool to community resource: What's the extra work and who will do it? *Proceedings of the Annual ACM Conference on Computer Supported Cooperative Work & Social Computing*, 18, 417–430.
- Uhlir, P. F., & Schröder, P. (2007). Open data for global science. *Data Science Journal*, 6(17), OD36–OD53.
- van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for values: An introduction. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.). *Handbook of ethics, values, and technological design: Sources, theory, values and application domains* (pp. 1–7). Dordrecht, Netherlands: Springer.
- Vayena, E., & Tasioulas, J. (2016). The dynamics of big data and human rights: The case of scientific research. *Philosophical Transactions of the Royal Society A*, 374(2083), 20160129.
- Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation*, 3(1), 114–126. <http://dx.doi.org/10.2218/ijdc.v3i1.46>.
- Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One*, 8(7), e67332. <http://dx.doi.org/10.1371/journal.pone.0067332>.
- Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, 33(5), 631–652. <http://dx.doi.org/10.1177/0162243907306704>.

Peter T. Darch is an assistant professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign. He received his PhD in computer science from the University of Oxford, England, in 2012. Prior to joining the iSchool at Illinois, Darch worked as a postdoctoral researcher in the UCLA Department of Information Studies. He is particularly interested in changes in the organization and conduct of contemporary scientific research that result from the interaction of technologies that afford collection of increasing quantities and types of scientific data with broader sociotechnical factors. To study these changes, he conducts longitudinal ethnographic studies of large, multidisciplinary scientific projects, applying theoretical approaches from science and technology studies to examine relationships between contexts in which these collaborations are embedded, information infrastructures, and scientists' day-to-day data practices. His work has been published across a range of journals, including the *International Journal on Digital Libraries* and *Philosophical Transactions A of the Royal Society*.

Emily J. M. Knox is an assistant professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign. Her research interests include information access, intellectual freedom and censorship, information ethics, information policy, and the intersection of print culture and reading practices. She recently edited *Trigger Warnings: History, Theory, Context* (Rowman & Littlefield, 2017). Her book, *Book Banning in 21st Century America* (Rowman & Littlefield, 2015) is the first monograph in the Beta Phi Mu Scholars' Series. Knox received her PhD from Rutgers University School of Communication and Information. She serves on the boards of the Association of Information Science and Technology, Freedom to Read Foundation and the National Coalition Against Censorship.