Prototype Ethics

Foundations for the Research Ethics of Real-World Technology Research

Mollen, J.K.

DOI
10.4233/uuid:725d6b2d-af31-4709-90e0-bbd5944aed03

Publication date
2025

Document Version
Final published version

Citation (APA)
Mollen, J. K. (2025). *Prototype Ethics: Foundations for the Research Ethics of Real-World Technology Research* . [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:725d6b2d-af31-4709-90e0-bbd5944aed03

Prototype Ethics

Foundations for the Research Ethics of Real-World Technology Research

Joost Mollen

Simon Stevin Series in the Ethics of Technology

Prototype Ethics

Prototype Ethics

Foundations for the Research Ethics of Real-World Technology Research

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. T.H.J.J. van der Hagen
Chair of the Board for Doctorates
to be defended publicly on
Wednesday 8, October 2025, at 12:30 o'clock

By

Joost Krijn MOLLEN
Master of Science in Media Technology
Leiden University, The Netherlands
born in Nijmegen, The Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. M.J. van den Hoven	Delft University of Technology, promotor
Dr. M.B.O.T. Klenk	Delft University of Technology, copromotor

Independent members:

Prof. dr. ir. I.R. van de Poel	Delft University of Technology
Prof. dr. T.A.P. Metze	Delft University of Technology
Prof. dr. E.M. van Bueren	Delft University of Technology
Prof. dr. L.E.M. Taylor	Tilburg University
Prof. dr. R. Hillerbrand	Karlsruhe Institute of Technology, Germany
Dr. ir. U. Pesch	Delft University of Technology, reserve member

Research for this thesis was made possible by the Province of South-Holland.

The Simon Stevin Series in the Ethics of Technology is an initiative of the 4TU Centre for Ethics and Technology. Contact: info@ethicsandtechnology.eu

ISBN: 978-94-6384-835-0
ISSN: 1574-941X

Copies of this publication may be ordered from the 4TU.Centre for Ethics and Technology, info@ethicsandtechnology.eu
For more information, see http://www.ethicsandtechnology.eu

Contents

Acknowledgements

Writing this doctoral thesis would not have been possible – or nearly as enjoyable – without the involvement of a great many people, my true gratitude to whom I can only approximate in written form.

First and foremost, I would like to thank my supervisors, Jeroen van den Hoven and Michael Klenk. They gave me the freedom to struggle and grow, held me to high standards, and were instrumental in my academic development. At times, they seemed to have more confidence in me than I had in myself, and I remain forever in their debt for their help in setting aside doubts and staying on course. I also want to specifically thank Michael for being a wonderful co-author.

This dissertation would not exist without the support of the Province of South-Holland, for which I'm immensely grateful. Thanks, everyone, for the warm welcome and specifically to Ivonne Jansen-Dings for being a great inspiration. I also want to thank Ibo van de Poel, Rafaela Hillerbrand, Linnet Taylor, Tamara Metze, Ellen van Bueren, and Udo Pesch for being on my defense committee.

I was humbled to be surrounded by an incredibly talented and kind group of colleagues at TU Delft, the Ethics/Philosophy of Technology section, the Delft Digital Ethics Centre, and the 4TU.Ethics community. Their feedback was of great benefit to my research, and it was a comfort to discuss shared challenges. I am also grateful to the supporting staff in our section for helping me navigate the intricacies of academic bureaucracy.

I want to thank my friends who — despite my protests — ripped me away from work and isolation and gave me a much-needed social life. My life would be incredibly dull without you all.

I want to thank my mam, pap, broertjes, zusje, Boomer, oma, and opa for their care and support during my whole academic trajectory. Specific thanks to oma for lighting candles and invoking the clearly present help of higher powers in bringing this doctoral thesis to a successful end.

Thanks to this thesis, I was lucky to make many wonderful new friends by attending inspiring conferences in beautiful places such as Lake Como, Vienna, Tokyo, and London. I'm especially grateful to this doctoral thesis for bringing my partner Stacy (and her cat Gibi) into my life. Your love, care, and wit were of great support when my boulder would roll down the mountain, and I'm beyond grateful to have you by my side.

Summary

Real-world technology research involves testing technologies in natural and uncontrolled environments that are (or resemble) the intervention's use setting. As research with technologies under real-world conditions has become a pervasive phenomenon in our public streets, homes, shops, jobs, and social media, scholars have drawn attention to its ethical concerns and the absence of research ethics governance, such as ethics guidelines and independent oversight. In the absence of research ethics for real-world research, scholars have evaluated various real-world research examples with existing research ethics principles and norms and have called for codes of ethics. However, this scholarship faces at least two shortcomings. First, this scholarly attention is fragmented across different disciplines, potentially at the expense of common ethical concerns and guardrails that should be applied to all real-world research formats, irrespective of their domain. Second, it is unclear whether existing research ethics norms – developed for (predominantly) controlled (human subject) scientific research – can capture the full range of ethical challenges shared by all research in the real world, potentially overlooking ethical concerns outside the proverbial lab. In this thesis, I provide a comprehensive analysis of real-world research to address these two gaps. I ask in which ways common ethical challenges emerge in research under real-world conditions, and in what ways research ethics principles and norms fail to account for these ethical challenges. I argue that real-world research shares common ethical salient characteristics, such as 'coupling' options to subjects, that need research ethics governance, but that the content of this research ethics governance cannot be wholly based on existing research ethics principles and norms. This is because real-world research raises novel ethical challenges, and existing research ethics norms, such as informed consent or the right to withdraw, cannot be upheld without severely altering the practice. This thesis, thus, lays the groundwork for an ethics of real-world research by developing its philosophical foundations and identifying common challenges to research under real-world conditions that such an ethics should take into account.

Samenvatting

In de afgelopen decennia worden nieuwe technologieën in toenemende mate getest in het dagelijks leven, zoals zelfrijdende auto's op de openbare weg en misdaadvoorspellende kunstmatige intelligentie in winkels of het nachtleven. Dergelijk veldonderzoek test nieuwe technologieën in natuurlijke omgevingen, zoals op straat, in winkels, bij mensen thuis of op online platforms, of in zogenaamde 'living labs' of 'fieldlabs', om zo representatieve en operationele kennis te vergaren. Technologische veldtesten worden gezien als een belangrijk instrument voor innovatiebeleid en voor adresseren van relevante maatschappelijke problemen. Maar ze kunnen ook ethische problemen veroorzaken. Zo kan de experimentele en technologische interventie in het dagelijkse leven van mensen kan gepaard gaan met risico's, bijvoorbeeld op het gebied van privacy, mensenrechtenschendingen en fysiek letsel. Desondanks ontbreken er in grote mate ethische richtlijnen en onafhankelijke toezichtmechanismen voor dit soort veldonderzoek.

In de academische literatuur wordt technologisch veldonderzoek op moreel en bestuurlijk vlak regelmatig geëvalueerd met bestaande en prominente (wetenschappelijke) onderzoeksethische kaders. Maar deze benadering kent minstens twee problemen. Ten eerste, aandacht voor de ethiek van technologische veldtesten is gefragmenteerd over verschillende academische disciplines en focust zich op verschillende ethische problemen en nieuwe technologieën. Deze fragmentatie staat een potentieel gemeenschappelijke benadering in de weg; een die gericht is op het identificeren en adresseren van gedeelde ethische uitdagingen en barrières die voor alle soorten veldtesten met nieuwe technologieën gelden, ongeacht het vakgebied. Ten tweede is het onduidelijk of prominente normen en principes in de onderzoeksethische literatuur – dat zich voornamelijk richt op gecontroleerd (mensgericht) wetenschappelijk onderzoek – het volle spectrum van ethische uitdagingen kan vangen dat technologische veldtesten met zich meebrengt. Hierdoor riskeren we dat we specifieke ethische uitdagingen van technologisch veldonderzoek over het hoofd zien.

In dit proefschrift adresseer ik deze twee problemen. Ik analyseer welke gemeenschappelijke ethische uitdagingen zich voordoen in technologisch veldonderzoek en op welke manieren de onderzoeksethiek tekortschiet in het vangen van deze uitdagingen. Ik beargumenteer dat technologische veldtesten, ongeacht hun focus of omgeving, gedeelde en ethisch relevante kenmerken hebben. Een voorbeeld hiervan

is dat technologisch veldonderzoek de opties van participanten op een unieke wijze aan elkaar koppelt. Daarnaast beargumenteer ik dat er ethische regulering voor technologische veldtesten nodig is, bijvoorbeeld in de vorm van ethische richtlijnen en extern toezicht, maar dat deze niet volledig gebaseerd kan worden op prominente onderzoeksethische principes en normen. Dit komt doordat een deel van deze normen zich richten op het beschermen van het individu, zoals kennisverschaffing, persoonlijke toestemming of het recht om zich terug te trekken zonder straf, maar dat het individu lastig te identificeren is in technologische veldexperimenten aangezien deze vorm van onderzoek intervenieert in openbare, ongecontroleerde testomgevingen.

Ik beargumenteer daarom dat we een nieuwe onderzoeksethiek zouden moeten ontwikkelen die rekening houdt met het specifieke karakter van technologisch veldonderzoek. Dit proefschrift legt de basis voor dit kader door de filosofische grondslagen ervan te ontwikkelen en gemeenschappelijke, ethische uitdagingen te identificeren waarmee de ethiek van technologisch veldonderzoek rekening zou moet houden.

1. Introduction

By going about our daily lives, we might find ourselves part of an experiment. In recent decades, the development of new technologies has become intimately entangled with many aspects of our lives, such as our public spaces, homes, online environments, and places of work and commerce. Examples of such real-world technological research include tests with self-driving cars on public roads (DeArman, 2019; Stilgoe, 2020), mood-altering street lights in nightlife streets (Galič, 2019), predictive policing technologies in public areas (Amnesty, 2020; Susser, 2021), migration management technologies at our borders (Molnar, 2020; Aradua, 2020), biometric technologies in humanitarian aid camps (Fejerskov, 2020), AI-driven technologies in active warzones (Hoijtink, 2022), experimental smart homes (Taylor, 2020), public interactive technologies (Waern, 2016), smart city interventions (Kitchen, 2016; Zimmerman, 2023), digital decentralized clinical trials (DCTs) at participants homes (Van Rijssel et al., 2022) and A/B tests on social media and online platforms (Kramer, 2012; Grimmelman, 2015; Polonioli et al., 2023; Rahman et al., 2023).

All these forms of research can differ in many aspects. They can differ in their subject, methodology, research environment, and actors involved. They might be referred to by various names – 'field experimentation' (McDermott & Hatemi, 2017), 'generative experiments' (Ansell & Bartenberger), 'action-guiding experiments' (Hansson, 2016), 'practical experiments' (Kroes, 2017), etc. They might be organized in so-called 'real-world testbeds' (Arntzen et al., 2019), 'living laboratories' (Baccarne et al., 2014), or 'real-world labs' (Dusseldorp, 2024); transdisciplinary collaborative spaces between academic, public, and private entities in which innovation can be promoted or the social value of new ideas and technologies explored, often in the absence or exemption of regulative demands (Engels et al., 2019; Ranchordas, 2021; Madiega & Van De Pol, 2022; Colonna, 2023).

However, despite their apparent differences, all these examples share that they involve researching interventions under so-called 'real-world' conditions — natural environments that are (or aim to resemble) the intervention's (eventual) use-setting. These research formats aim to bridge the gap between controlled laboratory conditions and the 'real world' by testing ideas and technologies in 'natural' and 'complex' environments (Artzen et al., 2019) and capturing (performance) data in representative real-world environments. The rationale is that while new ideas and innovations

might hold great potential to improve our lives, they need to be exposed to their complex 'real-world' use settings to evaluate their performance or impact in these environments and produce unique insights that are difficult or impossible to simulate or capture in more controlled research environments (Arntzen et al., 2019)[1]. In short, it is research aiming to capture real-world data in the real world, and as such, throughout this dissertation, I will refer to these research practices collectively as *real-world research*. They are the focus of this thesis.

Research under real-world conditions is an increasingly important instrument in innovation policy to develop and implement robust technologies (Arntzen, 2019; Engels et al., 2019). For example, real-world living labs and testbeds are employed to implement the United Nations Sustainable Development Goals (Molnar et al., 2023), promote the 'diffusion' of digital innovation (OECD, 2019), create opportunities for law enforcement agencies to test novel AI technologies (Europol, 2024), integrate smart city technologies in urban environments (Engels et al., 2019), can help identify harms to fundamental rights before AI models are fully publicly released (Janssen, 2020; Ministerie van Algemene Zaken, 2022) and contribute to ethical AI design, development, and deployment (Harbers & Overdiek, 2022). With the intent to promote innovation, specific real-world research, for example, with AI systems, is routinely enabled and encouraged by 'soft law' mechanisms such as regulatory sandboxes that provide developers the possibility to test innovations under the oversight of a regulatory body (Ranchordas, 2021; Madiega & Van De Pol, 2022) or made exempt from regulatory demands in AI governance regulations such as the European Union's AI Act (Colonna, 2023).

As real-world research has become a widespread strategy to address various societal and innovation challenges (Ansel & Bartenberger, 2016; 2017), this has prompted increased scholarly interest in real-world research from various perspectives, such as their need (Sherman, 2016), their legality (Ranchordas, 2021), their governance (Taylor, 2020; Stilgoe, 2020), their role in innovation governance (Renn, 2018; Engels et al., 2019), their politics (Evan & Karvonen, 2010; Evans et al., 2018; Hall & Hasan, 2020; Pfotenhauer, 2022), epistemology (Evan & Karvonen, 2010; Singer-Brodowski et al., 2018), human rights (Amnesty, 2020) and their ethics

[1] For various perspectives on the epistemology of real-world research practices, see, for example, Evans & Karvonen (2010), Ansell & Bartenberger (2016), Sherman et al. (2016), and Singer-Brodowski et al. (2018).

(Sainz, 2012; Kitchin, 2016; Taylor, 2020; Dusseldorp et al., 2024; Zimmerman, 2024).

This thesis is concerned with the ethics of real-world research.[2] Despite their potential practical and epistemic value, real-world research – as any research – can raise ethical concerns. Scholars have, for example, drawn attention to examples of real-world research exposing persons to risks (Maheshwari & Nyholm, 2022), harm (Stilgoe, 2020), manipulating populations (McDermott & Hatemi, 2020), violating their human rights (Amnesty, 2020), or subjecting persons to research against their knowledge or consent (Kitchin, 2016). By testing interventions under real-world conditions in a real-life environment, the people who live or frequent this physical or online environment are subjected to the research intervention and the (potential) associated risks, influences, and data capture, obfuscating a clear distinction between research participation and daily life. This can happen without their knowledge or consent, meaning people might end up as unknowing research subjects against their wishes. So, while real-world research might help develop robust new technologies, attention must be paid to conducting this research ethically. This is a matter of *research* ethics.

1.1. Research Ethics: Morality and Governance

In this thesis, two 'kinds' of research ethics are discussed, which need to be distinguished. First, research ethics, as a field of applied moral philosophy, is concerned with how research ought to be conducted, how researchers ought to behave, and how we ought to treat others in the name of research. It concerns moral judgment on the permissibility or acceptability of research and, for example, which principles and norms researchers should adhere to for their research to be ethical. In defining principles, I follow Beauchamp and Childress's claim that principles are "normative generalizations that guide actions" but which "leave considerable room for judgment in specific cases and that provide substantive guidance for the development of more detailed rules and policies" (Beauchamp & Childress, 1994, p. 38). Examples of such principles include 'non-maleficence,' 'beneficence,' 'justice,' and 'respect for personal autonomy' (Beauchamp & Childress, 1994). When referring to norms, I refer to

[2] While it predominantly focuses on examples of technological research – that is, research focused on the development of new (or improved) technologies, its findings are also relevant to real-world research in general.

these "more detailed rules" that researchers should follow (Beauchamp & Childress, 1994, p. 38). For example, while a more general principle might guide us to ensure that research respects a person's autonomy, a more detailed rule (norm) through which we might meet that principle could be to obtain the informed consent of persons before subjecting them to research.

One question that concerns the moral 'scope' of research ethics is to which researchers or research practices particular research ethics principles and norms (should) apply. Research ethics thought and literature focus predominantly on the moral evaluation of scientific research. However, much real-world research is not scientific research but instead involves research focused on the acquisition of practical, trial-and-error, local (and thus potentially less generalizable) insights. Such practices prompt the question of whether such practices ought (as a matter of morality) to follow similar moral principles and norms to scientific research. For example, researchers working at corporations or city governments might, in an institutional or legal sense, not be required to adhere to particular principles and norms (for example, a norm to obtain informed consent or provide subjects with a right to withdraw from research without penalty).[3] However, should they, as a matter of morality, uphold such principles and norms nonetheless? Should these principles and norms be the same as their scientific counterparts? Or might we have good reasons to treat these parties with a distinct ethical approach? Such questions concern the moral scope of research ethics.

Second, as a field of governance, research ethics is concerned with regulating research through policies, laws, protocols, guidelines, committees, and enforcement structures to ensure research is conducted in compliance with, for example, ethical guidelines or codes, the judgments of ethical review boards, and legal requirements (Kolstoe & Pugh, 2024). Examples of ethical guidelines or codes include the Nuremberg Code, Helsinki Declaration, CIOMS Ethical Guidelines, and the Belmont Report, outlining principles or norms that researchers should comply with for their research to be ethical. The above-mentioned moral principles, for example, have influenced much of contemporary research ethics governance.[4]

---

[3] Assuming there is no law or internal institutional guideline that prescribes this.

[4] These principles also influenced AI ethics guidelines such as the EU's Ethics Guidelines for Trustworthy Artificial Intelligence or OECD's Recommendation of the Council on Artificial Intelligence (Nikolinakos, 2023; Porter et al., 2024).

Unlike the moral scope of research ethics, the scope of research ethics in an institutional or legal sense (e.g., its governance) extends only to all the researchers and research practices within an institution's domain of authority or the law's jurisdiction. For example, I fall within the scope of the research ethics governance of my university, which means I have to comply with particular research ethics requirements that my university prescribes, such as following particular protocols (e.g., submitting a research proposal for independent review when human subjects are involved) and follow particular ethical norms (e.g., taking precautions to mitigate harm to subjects in my research). At the same time, additional conditions limit the institutional scope of research ethics governance within my institution. For example, when my research does not involve human subjects and only involves desk research (such as this dissertation), per the rules outlined by my university, I do not need approval from our research ethics committee. The rules at another university might differ, but they do not apply to me since I operate outside its institutional scope. The legal scope of research ethics works much the same, yet now concerns whether I, as a researcher, or my research project, is subject to a particular law regulating research conduct. Similarly, its scope is bound by the practices and persons it targets and its jurisdiction. For example, a law dictating how biomedical research in the Netherlands ought to be conducted does not affect my philosophical research in the Netherlands. It would neither affect a biomedical research project in Canada, seeing how it is conducted outside the jurisdiction of Dutch law.

Thus, claims about whether particular research ethics principles and norms (should) apply to particular real-world research types can be claims in a moral sense (i.e., 'due to moral reasons, the moral norms of X also apply to Y' (and we did not realize or were mistaken before)) or in an institutional or legal sense (i.e., 'those practices fall under a particular authority and we should govern them as such,' 'we should change the rules so that those practices fall within the scope of jurisdiction of the law or authority of an institution' or 'we should develop new institutions to enforce particular principles and norms,' etc.) These claims can also overlap. Arguments advocating for changing the regulation or the scope of research ethics governance might rest on a moral argument. For example, one could argue that principles or norms apply to specific researchers or a new type of research as a matter of morality and that we should, thus, change our regulations to reflect this. Grimmelman has, for example, made such a claim, writing:

> "To the extent that the Common Rule [*red.* a United States legal rule of ethics regarding biomedical and behavioral human subject research] reflects a consensus about academic research on social media users, it should also extend to corporate research on social media users because the ethical argument for regulating the latter is at least as strong as the argument for regulating the former" (Grimmelman, 2015, p. 254).

It is important to differentiate these two sides of research ethics. This is because, while ideally, these 'sides' of research ethics overlap, meaning that research ethics governance is grounded on solid philosophical and moral foundations, they can diverge. We might, for example, overregulate particular research interventions and have no good moral reasons to do so, and vice versa. Alternatively, people might hold the belief that particular research ethics principles or norms do or do not apply as a matter of morality, yet turn out to be mistaken in their belief or overlook relevant philosophical concepts in their moral evaluation of research. A seeming disconnect between these two sides of research ethics in the context of real-world research is a central focus of this dissertation.

1.2. A Lack of Research Ethics Governance in the Wild

Scholars have increasingly pointed out that real-world research lacks research ethics governance – institutional guidelines, committees, protocols, or laws – to address or help mitigate its research ethical challenges.[5] For example, commenting on corporate experimentation on online worker platforms in the Financial Times, Rahman writes:

> "The problem is not experimentation in itself, which can be useful to help companies make data-driven decisions. It is that most do not have any internal or external mechanisms to ensure that experiments are clearly beneficial to their users, as well as themselves. Countries also lack strong regulatory frameworks to govern how organizations use online experiments and the spillover effects they can have. Without guardrails, the consequences of unregulated experimentation can be disastrous for everyone" (Rahman, 2024).

[5] Note that this is different from accounts that argue that particular research interventions are subject to norms in an institutional or legal sense, yet they are not followed by researchers. For example, McDermott and Hatemi have noted that social scientists routinely do not comply with the research ethics guidelines that apply in their field (2020).

In the context of testing new technologies for migration management in border zones, Molnar writes that:

> "All this experimentation occurs in a space that is largely unregulated, with weak oversight and governance mechanisms, driven by the private sector innovation" (2020, p. 34).

Additionally, in the context of urban experimentation, Taylor argues that:

> "What is missing so far, however, is an interrogation of urban experimentation that takes seriously the issue of research on human subjects, and asks what norms, rules and boundaries are appropriate" (Taylor, 2020, p.1903).

This absence is particularly pronounced in contrast to scientific research and other publicly-funded research in which research ethics governance is firmly established. For example, in the context of online corporate A/B testing, Polonioli and colleagues argue that:

> "The use of human subjects in research that is not federally or publicly funded—such as in the case of privately funded A/B testing, often affecting millions of potentially unaware people—has remained unregulated" (Polonioli et al., 2023, p. 669).

Additionally, Weiss writes that:

> "Social scientists follow strict Institutional Review Board (IRB) procedures that govern the ethics of experiments involving people — such as informing them and requiring consent — but these rules do not apply to technology companies. And that is leading to questionable practices and potentially unreliable results … Technology companies use their terms of service to authorize them to collect data without any obligation to inform people that they were involved or provide any opportunity for them to withdraw. Thus, digital experimentation faces scant oversight" (Weiss, 2024, p.259).

Alternatively, Calo writes that:

> "Any academic researcher who would conduct experiments involving people is obligated to comply with robust ethical principles and guidelines for the protection of human subjects, even if the purpose of the experiment is to benefit those people or society… But a private company that would conduct experiments involving thousands of consumers using the same basic techniques, facilities, and personnel faces no such obligations, even where the purpose is to profit at the expense of the research subject" (Calo, 2013, p.101).

While these accounts target different research domains, methodologies, and parties, they all highlight a common problem: a gap in research ethics governance regarding

research *in the wild*. This gap is particularly pronounced when compared to the research ethics governance of scientific research, revealing a seeming unequal treatment regarding what institutional demands we place on researchers' conduct and the protections we offer for research subjects between research domains. I take a descriptive stance regarding whether these examples are, in fact, not subject to research ethics governance in an institutional or legal sense by drawing from claims in the literature that support this. Normatively, I argue that this absence is important since research ethics governance, at least theoretically, aims to provide guardrails to protect persons and society from research misconduct and harm. Thus, I will argue that the absence of research ethics governance for real-world research comes at the potential cost of those the research potentially affects, which ought to be remedied.

1.3. Addressing the Gap and Unifying a Fragmented Debate

In the absence of research ethics governance for real-world research, such as specific ethical guidelines, scholars have taken existing research ethical guidelines from other research domains (and the principles and norms outlined within them) to analyze and evaluate various real-world research practices and discussed how such guidelines can be adapted for particular real-world research domains. For example, Svensson and Hansson have taken ethical principles prominent within biomedical research and used this frame to analyze traffic experiments, noting relevant differences and similarities between the two research practices and how the ethical principles of biomedical research could be adapted for traffic research (Svensson & Hansson, 2007). Similarly, Zimmerman has taken ethical guidelines from the psychological sciences and discusses their applicability to smart city experiments (Zimmermann, 2023). Alternatively, Benbunan-Fich has analyzed various examples of corporate social media A/B testing and concluded that these experiments did not comply with existing research ethics guidelines for the treatment of research subjects, noting that companies are not bound by law to do so and recommending the development of a new ‘code of ethics’ for online research, ‘safeguard mechanisms’ and independent review and regulation (Benbunan-Fich, 2017).

However, what, if anything, makes existing research ethics frameworks an appropriate tool through which to analyze or evaluate real-world research? Why should we analyze and evaluate corporate or government researchers with the same

research ethical principles and norms as their scientific counterparts?[6] This has not been adequately explored. As a field of inquiry, research ethics has predominantly focused on ethical challenges for *scientific* (human subject) research, reflecting its current regulative scope. Research ethics emerged as what London calls a "practical policy response to revelations of abuse," predominantly by biomedical and behavioral researchers (2020, p.27). Hence, the earliest and most influential research ethics guidelines focused on ethical questions and challenges within these domains. While these challenges might overlap with real-world research practices, they do not necessarily need to.

There are good reasons to believe this overlap is indeed limited. Scholars have, for example, drawn attention to how existing research ethics guidelines, specifically the guidelines for human subject research, are poorly equipped to deal with domains such as big data and AI research (Metcalf & Crawford, 2016; Ada Lovelace Institute, 2022), political field experiments (Whitfield, 2019; McDermott & Hatemi, 2020), and cluster randomized trials (Weijer et al., 2011; 2012). For example, Whitfield has argued for a separate ethics of political field experiments because, he argues, these experiments bring about unique ethical challenges that existing research ethics guidelines do not capture (2019). Other accounts have argued that particular real-world research practices often present a mismatch due to the focus of existing research ethics norms on the research conducted on the individual: for example, data subjects do not neatly fit into the definition of human subjects (Metcalf & Crawford, 2016), and cluster randomized trials can take entire groups or populations as their unit of intervention rather than individuals (Weijer et al., 2011). Such accounts show that existing research ethics guidelines are not necessarily a catch-all for all research practices.

Thus, even if we agree that real-world research should at least be subject to some form of research ethics governance[7], existing research ethics principles and norms from other research domains might not necessarily offer the correct set of normative content to capture and govern the ethical challenges of real-world research. Real-world research might raise ethical challenges that are not found in 'the lab.' Addi-

---

6 If one were to believe that different types of research should be subjected to different types of moral obligations – as, for example, Meyer has argued in the case of corporate A/B testing (2015) – then most of the concerns pointed out by some accounts are rendered moot.

7 Some scholars have called for the development of ethical protocols or codes of conduct for real-world research, for example, such as A/B Testing (Benbunan-Fich, 2016; Weiss, 2024).

tionally, there might be meaningful differences between these two research domains[8] that warrant a different ethical approach. Addressing this concern is important because if we analyze and evaluate real-world research with imperfect frameworks, we risk a disproportionate focus on norms that cannot be upheld in real-world settings (e.g., individual informed consent) or overlook particular moral salient concerns that might hold for all relevant examples of real-world research.

Concerning the latter point, the scholarship on the ethics of real-world research has remained scattered across different debates and research domains. Scholars have focused on domain-specific issues such as specific technologies (e.g., self-driving cars, smart city applications, etc.), ethical concerns (e.g., privacy), or fields of research (e.g., data science, traffic research, etc.). This fragmentation of the debate is not surprising, considering the vivid contrasts between examples of research conducted under real-world conditions. For example, testing self-driving cars on public roads seems in stark contrast with A/B testing on social media platforms, and their localized ethical challenges might be very different. However, the lack of research ethics governance is a common problem for many relevant real-world research examples, revealing a need for a more comprehensive approach to the ethics of real-world research.

This thesis will offer and defend such a comprehensive account of the ethics of real-world research. By this comprehensive approach, I mean an account of the ethics of real-world research that goes beyond the particulars of individual examples involved and instead focuses on (ethical) challenges brought about by the characteristics of the real-world research itself. Such an approach is important because a fragmented debate might overlook a deeper ethical concern that unifies these seemingly varied and diverse new forms of research. For example, domain-specific accounts such as Whitfield's argument about the separate ethics of political field experiments are limited to political science since it is grounded on ethical challenges that specifically political field experiments bring about (2019). However, attention to domain-specific issues may come at the expense of finding guardrails that should be applied to all real-world research formats, notwithstanding their domain. As a consequence, our moral and regulatory response may fall short.

Consequently, this thesis argues that we must realize a research ethics governance based on ethical content that addresses the shared ethical challenges of real-

[8] For example, Svensson and Hansson (2007) discuss relevant differences between biomedical and traffic research in the context of research ethics.

world research. This thesis argues that the ethical content of this governance cannot be wholly the same as outlined in existing research ethics guidelines. I argue that a research ethics governance of real-world research would face serious obstacles if they used the ethical content of existing research ethics principles and norms because they do not capture all that might be wrong within real-world research and feature norms that are impossible or difficult to uphold in real-world research practice due to the lack of control over the research environment in real-world research.

My analysis shows that many relevant real-world research examples would conflict with these principles and norms if we use existing research ethics guidelines to analyze real-world research. The upshot of this outcome would be twofold: either to find overriding reasons for all norms that are impossible to uphold in the real world (the obstacles to this approach are outlined in Chapter 4), or we have to amend real-world research practice in order so that they can uphold these norms, and not allow those that cannot do this. This latter point would be a problem considering the belief that real-world research offers important practical and epistemic benefits to develop new and robust innovations that other methodologies do not. A better and third solution, I argue throughout this dissertation, is not to wholly use existing ethics guidelines and instead develop research ethics governance for real-world research that accounts for the specific ethical challenges of real-world research and contains ethical content that aligns with the realities of real-world research practice.

This thesis thus lays the groundwork for the ethics of real-world research. It develops its philosophical foundations and identifies challenges and areas of attention that such a framework should consider. However, this thesis does not develop such a framework. I argue that before we can develop such a framework, particular groundwork is necessary, which this thesis helps to address. Notably, a clear case needs to be made that (1) we can consider real-world research a unified and morally interesting unit of analysis (Chapter 2), (2) there is a need for research ethics governance (Chapter 3), in which ways common ethical challenges emerge in research under real-world conditions (Chapter 2 and 4), and (3) what ways do existing research ethics frameworks fail to account for these ethical challenges and thus need to be accounted for in a new research ethics framework (Chapter 4,5 and 6). I will outline the thesis in more detail in the following three sections. First, I will discuss the research questions, second, the methodology, and third, the scope.

1.4. Thesis Overview and Research Questions

In this section, I present a more detailed overview of this thesis and its research questions. The gap outlined in the previous section reveals a need for a comprehensive analysis of (the challenges of a) research ethics of real-world research. Thus, the main research question that this thesis aims to answer is:

> RQ: *In which ways do common ethical challenges emerge in research under real-world conditions, and in what ways do research ethics principles and norms fail to account for these ethical challenges?*

One challenge to such a unified approach — and an explanation for why the field is so fragmented — is that examples of research conducted under real-world conditions can vary wildly depending on the context specifics involved. They can differ in research environments, aims, methodologies, agents involved, and the domain about which knowledge is sought. For example, testing self-driving cars on public roads contrasts with A/B testing on social media platforms. This apparent diversity makes it unclear whether real-world research has unique, morally unifying characteristics that could justify a comprehensive approach to the ethics of real-world research. Hence, in Chapter 2, I lay the groundwork for a comprehensive ethical analysis of real-world research. Specifically, I address the question:

> RQ2: *Are there unifying and ethically significant features common to all relevant examples of real-world research*?

I identify such a feature and, in doing so, justify the focus of this thesis as a legitimate unit of research and ethically interesting phenomenon. That feature is 'coupling.' With coupling, I offer a novel account of what unifies seemingly diverse forms of real-world research, ranging from self-driving car tests to online A/B tests, and which can be considered a morally salient feature. I argue that real-world research has a seemingly unique quality of being able to 'couple' itself intimately with the daily lives of persons at places of work, on public streets, and on our social media. Coupling occurs, for example, when research is conducted under real-world conditions in a public street, and a person can no longer engage with this environment –

walk that street — without being subjected to the research.[9] Before the research was set up, people had the option of 'going to the street' without the option of 'being a subject.' After the experiment is set up, both options are coupled, and one can no longer accept or reject one without the other. Coupling is thus defined as when two potentially independent options are 'coupled,' meaning you cannot choose (or reject) one without the other.

I argue that this coupling can give a reason for moral concern, depending on the moral salience of (and the interplay between) the options themselves that are coupled, as well as the degree of control one has over the coupling in the first place. In other words, the 'moral weight' of the coupled options that one has to accept or reject gives moral salience to the dynamic of coupling. This account suggests that we should view real-world research that couples weighty, significant choices, like where one lives with an experiment, with more suspicion than research that couples more insignificant choices, such as which street one visits in a city. Additionally, the more negative moral salience the experimental option presents, the more we should view the resulting real-world research with suspicion.

Having offered a clear description of real-world research and identified a justified reason for a unified approach to the ethics of real-world research, in Chapter 3, I discuss the general absence of ethical governance in real-world research. Real-world research practices are not held to institutional or legal ethical standards in the same way that scientific research might. However, current scholarship has been fragmented in addressing this issue. So, in Chapter 3, I continue my comprehensive approach and aim to analyze potential reasons for this absence of real-world research. Specifically, in Chapter 3, I aim to answer the following question:

> RQ3: *What, if anything, justifies the lack of research ethics governance for real-world research in comparison to scientific research?*

I will argue that this absence is unjustified and that we have good reasons to develop research ethics governance for real-world research. However, this does not necessarily mean that this research ethics governance must be based on existing and

---

[9] Throughout this thesis, I use (research) subjects rather than (research) participants. This terminology does not assume that they voluntarily engaged, are informed, etc. They are subject solely in virtue of being *subject to* the research, either being causally influenced by it or being observed in some sense, e.g. through data gathering.

prominent research ethics norms and principles. Instead, I examine the current absence of ethical governance and reject four possible justificatory reasons that might justify this current differential treatment (i.e., placing different regulatory demands on various types of research). These are (1) the ability or potential to be harmful, (2) their environmental research conditions, (3) the goal of the research, and (4) the role of the researcher. Additionally, I argue that asymmetrical research ethics regulatory demands create the opportunity for the potential avoidance of research ethics burdens by placing research activities outside the current scope of research ethics regulation, which comes at the expense of those whom these demands aim to help protect.

After having established in Chapter 2 that we have good reasons for a unified ethical approach to real-world research and in Chapter 3 that there are good reasons to develop research ethics governance to real-world research in general, in Chapter 4, I examine to what extent such a research ethics governance of real-world research can draw from paradigmatic research ethics norms. With paradigmatic research ethics norms, I mean those norms that are prominent in the literature and ethical guidelines. To clarify, I do not presuppose that the moral obligations of researchers conducting real-world research should necessarily be the same as paradigmatic ethical norms in standard human research experiments as a matter of morality. Instead, I note that these paradigmatic norms have been used to evaluate relevant examples of real-world research and ask – if we were to hold real-world research to these norms – whether such norms can be upheld in real-world research practice. Specifically, I ask:

> RQ4: *What challenges emerge in applying paradigmatic research ethics norms to real-world research?*

In Chapter 4, I identify and focus on one such challenge. I argue that real-world research presents a common problem to paradigmatic research ethics norms. Real-world research brings about a problem of identification. Research formats conducted under real-world conditions often involve significant uncertainty and difficulty in identifying their exact scope or reach. However, many paradigmatic research ethics norms require an investigator to know the identity of participants, for example, since they focus on protecting individuals. This means that many established research ethics norms – such as providing information about the research, informed consent,

and a just distribution of research participants – are difficult or even impossible to comply with in a real-world setting.

I refer to this problem as the identification problem of real-world research. While different authors have touched upon this identification problem, it has not been sufficiently explored, and its implications have been underestimated. Authors discussing the problem have generally focused on informed consent and neglected to discuss other research ethics norms that rest on a similar assumption of participant identification.

I argue that the identification problem has far-reaching implications for our efforts to conduct research under real-world conditions since it would mean that we cannot neatly apply many paradigmatic research ethics norms to real-world research. The analysis shows that if we were to hold real-world research to these norms – as some scholars have done – then this renders much of real-world research unethical if overridden reasons cannot be found. So, while this thesis argues that real-world research practices should be subject to research ethics governance, in Chapter 4, I show that if we base the ethical content of this governance on paradigmatic research ethics norms, this would require severe research redesigns to ensure such norms can be upheld, which might impact their epistemic value.

In the previous three chapters, my analysis of the ethics of real-world research has been focused on real-world research at large. That is, I focused on challenges and issues that concern all examples of real-world research: a unifying moral concern (Chapter 2), a lack of ethical governance (Chapter 3), and a shared identification problem with the application of paradigmatic research ethics norms (Chapter 4).

In the final two chapters of the thesis, I further investigate this last issue by analyzing two distinct case studies of real-world research: real-world AI research (Chapter 5) and live-in laboratories (Chapter 6). In doing so, I aim to accomplish two goals. First, I will show how the themes raised in this dissertation play out on a concrete case basis. In doing so, these chapters address specific practitioner audiences, respectively, the generative AI community and the live-in laboratory community, raising awareness about research ethics challenges of real-world research within those research communities. Second, I aim to show that even though, at face value, these examples of real-world research have stark contrasts, they both present similar ethical concerns, which are grounded in them being real-world research.

First, in Chapter 5, I analyze the ethical challenges of real-world research with LLMs and generative AI. Specifically, I ask:

RQ5: *What challenges arise when we evaluate real-world AI research with paradigmatic research ethics principles?*

I argue that real-world AI research faces challenges in meeting paradigmatic research ethics principles — non-maleficence, beneficence, respect for autonomy, and distributive justice — for reasons that are grounded in them being operated in the 'real world' and that the absent or imperfect current research ethics governance of this real-world research exacerbates these challenges.

These challenges are important to address since real-world AI research presents ethical concerns, such as the potential for interpersonal and societal research harms, the increased privatization of AI learning, and the unjust distribution of benefits and risks. However, I discuss these ethical concerns in the context of the epistemic need for real-world research with large language models. I discuss alternatives to real-world research, such as controlled or anticipatory learning methods such as laboratory benchmarking and forecasting, which have limitations exacerbated by large-language models' opaque internal operations and potential for emergent behavior. This epistemic need for real-world research – at least in the case of generative AI – makes these ethical concerns more pressing because if we halted or altered the practice, we might miss out on its potential epistemic benefit.

In Chapter 6, I continue my focus on the ethics of real-world research on a case-study basis by turning to live-in laboratories. These are (smart) homes built as experimental living environments to test the performance of novel technologies on their residents. Two important themes come together in this example. First, a significant theme in this thesis (especially Chapter 2) is that real-world research 'couples' itself with daily life. This theme is perhaps most pronounced in live-in laboratories, which 'couple' in a very clear sense with the daily life of research participants. Residents live in an experiment, and they would have to move house to no longer be research participants. Second, another theme is that paradigmatic research ethics norms are difficult to uphold when applied to real-world research. I show how these two themes come together in live-in laboratories by analyzing how the 'right to withdraw' — a paradigmatic research ethics norm that grants research subjects the ability to withdraw from research without penalty or coercive influences in order to safeguard the voluntary status of research participation – conflicts with this real-world research practice. Specifically, I ask:

RQ6: *How, if at all, does the right to withdraw conflict with live-in laboratory research?*

I argue that live-in laboratory research conflicts with this paradigmatic norm. I argue that since withdrawing from the live-in laboratory as a participant's primary residence means losing one's home, this creates negative financial and psychological consequences for participants. Such costs conflict with a participant's right to withdraw on two counts. First, the exit costs from the live-in laboratory constitute a penalty, and second, the costs of withdrawing from the live-in laboratory function as a constraint on a participant's liberty. If we were to take the right to withdraw seriously in this context, then coupling a participant's primary residence to research participation would be ethically problematic.

Finally, in my Conclusion, I summarize my argument and main points, highlight the implications of my argument, and look ahead and discuss outstanding questions and promising avenues for further research. To reiterate, the core claims of this thesis are that (1) real-world research warrants a comprehensive ethical approach because (2) they are unified by a morally salient characteristic, namely that they couple options for subjects. At the same time (3), real-world research lacks research ethical governance. I argue that (4) we should develop such a research ethics governance for real-world research, (5) but that the content of this research ethics governance cannot draw whole from paradigmatic research ethics principles norms. This is because (6) the characteristics of real-world research conflict with assumptions underpinning particular research ethics principles norms, and due to low control over the research environment of real-world research, this makes upholding a wide range of established norms difficult. Thus, in moving towards a research ethics of real-world research, we need to realize a new set of research ethics norms sensitive to its shared characteristics and challenges.

1.5. Methodology

In this dissertation, I will rely predominantly on philosophical and qualitative research; both in the literature I draw from and my methodology. Specifically, I take an applied philosophical approach in that I apply philosophical thought to real-world scenarios. Specifically, this thesis draws from research ethics, which, as a field of applied moral philosophy, is concerned with how research ought to be conducted, how researchers ought to behave, and how we ought to treat others in the name of research. This is an apt methodology since the focus of this thesis is both a practical concern, namely the ethics of research with technology under real-world conditions,

and concerns philosophical questions regarding how such research should be conducted and governed.

Additionally, I will supplement my philosophical analysis with insights and use cases from various academic fields that have discussed the ethical challenges of conducting research under real-world conditions within their domain. Examples include digital ethics, technology ethics, political science, social science, AI and data science, innovation studies, science and technology studies, and innovation studies. I do so for three reasons. First, to build a comprehensive account of real-world research, my account needs to include (and be able to account for) a wide variety of real-world research examples. Second, by drawing from various disciplines, I position this thesis as part of a more extensive debate spanning various research communities about the shortcomings of the scope and content of research ethics. Third, my analysis aims to bridge the theoretical and the practical and to make an impact beyond the fields of research ethics and technology ethics and also be of interest to a wide variety of scholars, practitioners from specific technological fields, and government facilitators interested in (the research ethics and governance of) research under real-world conditions. While this thesis draws inspiration from real-world data points, this dissertation does not conduct empirical research.

Finally, in recent years, there has been increased scholarly attention to the challenges of international research ethics (Benatar & Singer, 2000; Benatar, 2004; Schücklenk & Ashcroft, 2010) and whether (or not) research ethics should be considered 'universal' (Zhang, 2016; Msoroka & Amundsen, 2017). In this light, it is prudent to reflect briefly on this thesis and it's examples from that context. A majority of cases and literature selected within this thesis are from what can generally be referred to as the global north or generally Western countries (for examples of real-world research in the Global South, see Fejerskov 2020). The author of this thesis is also positioned within this context, and I am aware that such positionality could potentially limit the analysis presented in this thesis and give rise to particular blind spots. However, since this thesis largely focuses on identifying common (ethical) challenges of real-world research that supersede the specifics of the real-world context in which they are conducted, this positionality is considered to have limited influence on the findings of this thesis. Hence, discussions about the universal nature of research ethics remain outside the scope of this thesis. Nevertheless, future research on how various local contexts might give rise to unique real-world research ethical questions would be valuable.

1.6. Limiting the Scope

Here, I want to briefly clarify the scope of this thesis by distinguishing between the use of 'real-world research' in this thesis and, first, the idea of technology as a social experiment within sociology and the philosophy of technology and, second, the concept of 'experimentality' within anthropology and science and technology studies.

First, in recent decades, sociologists and philosophers of science and technology have increasingly conceptualized technology (or its introduction in society) as an experiment of sorts, for example, by framing society as a 'laboratory' (Krohn & Weyer, 1994), or the introduction of technology as a 'real-world experiment' (Gross, 2018; David & Gross, 2019), 'social experiment' (Martin & Schinzinger, 1983; Van de Poel, 2013; 2016; 2017) or a 'collective experiment' (Latour, 2004; Stilgoe, 2016). Notwithstanding the value of this frame, the idea that the introduction of technology should be thought of as a social experiment and the real-world research that this thesis discusses, while apparently similar, refer to two distinct phenomena.

The intention of the idea of technology as a social experiment is to draw attention to the inherent uncertainty involved in the introduction of experimental technologies in society (Stilgoe, 2016). Additionally, the frame of 'technology as a social experiment' is used to answer questions about the moral acceptability of new technologies as if we were to evaluate them on their status as experiments, i.e., were we to conceive of this technology introduction into society as an experiment, then would this be an ethical experiment? (Van de Poel, 2013). Moreover, the concept of experimentation is used to stress that we *should* learn from this uncertainty (Van de Poel, 2013; Stilgoe, 2016). In essence, it is an inherently normative concept[10].

Second, there is the concept of 'experimentality' in anthropology and science and technology studies (STS) (Petryna, 2007; 2009). Petryna uses this concept to draw attention to how many experimental drug trials replace local healthcare access, effectively becoming 'governance tools' for distributing public health resources. In a broader sense, the concept expresses the idea of 'government-by-experiment' (Hoijtink, 2022) and making societal challenges 'governable through experiments'

[10] These accounts have also drawn criticism, for example, that the frame of 'technology as social experiment' is essentially irrelevant (Peterson, 2013; 2017), that these experiments are different from experiments in the natural and social sciences (Kroes, 2015), or that the concept of experimentation is stretched out to a point where it loses analytical value (Karvonen & van Heur, 2014; Huitema et al., 2018; Hansson, 2019, notes 1).

(Aradua, 2022). For example, Fejerskov describes an 'experimental movement' made up of various actors active in the Global South that tackle social problems through technological innovation, for example, humanitarian organizations using experimental biometric technologies to identify and deliver aid to refugees (2022).[11] Experimentality has also been applied to cases like innovation in disaster relief (Hunt, 2019), technology-mediated treatment of migrants in border zones (Aradua, 2020), experimental innovation of warfare (Hoijtink, 2022), and the global humanitarian HIV programs (Nguyen, 2009). The concept also highlights how these environments' exceptionality often legitimizes experimentation. For example, Nguyen writes how HIV programs have become "epistemological practices that are qualified as *exceptional*, organized as an *experiment* and legitimated by the 'emergency' that requires immediate intervention" (2009, p. 211). In short, experimentality describes an approach and understanding of (innovation) governance that focuses on in-situ prototyping and trial-and-error learning to address social issues.

Naturally, there can be an overlap between the three ideas in various ways. For example, (emerging) technologies could be introduced into society through real-world research (e.g., Van de Poel's idea of 'learning-by-experimentation' (2017) and going from tacit to explicit social experimentation' refers to this idea (2017b)). Second, if we conceive of technology as a social experiment, we might end up evaluating this introduction with similar moral standards from research ethics as we do real-world research. This is, for example, what Van de Poel has done in developing an ethical framework for evaluating experimental technology (which he defines as technology about which we have "limited operational experience") when released into society (Van de Poel, 2016, p. 669). Furthermore, many real-world research might be driven by a rationale of 'experimentality.'

However, when I refer to real-world research throughout this thesis — or occasionally to real-world experimentation when such research concerns active intervention rather than observation — I refer to an explicit and concrete set of research formats that share at least two features. First, they are all *epistemic* activities. They explicitly intend to learn something about the intervention. Second, they are conducted under real-world conditions. In contrast to laboratory research or controlled experiments, all research under real-world conditions is characterized by relaxed, lower degrees, or even an absence of controls by the researcher on the

[11] For a review of Fejerskov's book, see Mollen, J. (2023). A. Fejerskov, The Global Lab: Inequality, Technology, and the Experimental Movement. *Prometheus, 39*(3), 189-194.

environmental research conditions (Harrisson, 2005; Ansell & Bartenberger, 2017; Kroes, 2016). Here, control refers to the control investigators have over the experimental intervention and environment to preserve the epistemic validity of a particular research outcome, such as through mechanisms such as randomization, control groups, isolating variables, or statistics (Ansell & Bartenberger, 2017). The absence of this control ensures the 'uncontrolled' complexity of a particular environment in which the research project takes place. However, it also has consequences for real-world research to uphold particular research ethics norms, as I will outline in Chapter 4. First, in Chapter 2, I will further clarify the concept of real-world research and argue that despite their apparent variety, they are unified in a common moral concern.

Prologue to Chapter 2

In Chapter 2, I defend a comprehensive approach to the ethics of real-world research with emerging technologies. This chapter[12] was co-authored with Michael Klenk. Prompted by the wide variety of examples of research conducted under real-world conditions – ranging, for example, from self-driving car tests on public roads to A/B testing on social media platforms – we asked whether these starkly different examples are, in some sense, morally unified, in that they share a common ethical concern. Of course, as mentioned in the introduction, all these examples are unified because they are research practices that happen under real-world conditions. However, as we will point out, this seems hardly a unifying *moral* feature. Hence, I aim to find a morally salient feature that unites all these examples. Specifically, I aim to answer the following question:

> RQ2: *Are there unifying and ethically significant features common to all relevant examples of real-world research*?

Answering this question is important because it provides the analytical foundations of this thesis. It justifies real-world research as a legitimate unit of philosophical research *and* as an ethically interesting phenomenon. I will argue that, in an important sense, all relevant examples of real-world research are unified and, in doing so, justify the focus of this thesis as a legitimate unit of research and an ethically interesting phenomenon. I provide a novel account of what unifies seemingly diverse forms of real-world research, ranging from self-driving car tests to online A/B tests, and can be considered a morally salient feature. That feature is 'coupling.'

---

[12] Submitted as 'Entangled Experiments: Coupling and the Ethics of Real-World Research' by Joost Mollen and Michael Klenk.

2. Coupling and Real-World Research

Abstract

Research conducted under real-world conditions, such as testing self-driving cars in public or A/B testing on websites, is an increasingly pervasive phenomenon. However, ethical inquiry into this phenomenon has been piecemeal and focused on domain-specific ethical concerns. In this paper, we argue that real-world research is unified as a phenomenon by a common and morally salient feature: option 'coupling.' By testing interventions in a particular 'real-world' environment, people can no longer engage with that environment without being subjected to the research's risks, influences, and data capture – thus 'coupling' these options. We identify coupling in a range of interesting examples of real-world research and point out grounds for moral concern, thus paving the way for a comprehensive approach to the ethics of real-world research.

2.1. Introduction

Our world is complex and complicated. So, to develop new policies and technologies that work, we must test how such innovations operate vis-à-vis the world's complexity. This practical and epistemic need has led to an increasingly pervasive phenomenon: research conducted under what is often called 'real-world' conditions as opposed to controlled research environments such as the laboratory (Ansel & Bartenberger, 2017). Examples include tests in public and private spaces with emerging technologies such as self-driving cars (DeArman, 2019; Stilgoe, 2020) or predictive policing technologies (Galič,2019; Amnesty, 2020; Susser, 2021), online A/B tests on social media or digital work platforms (Kramer, 2012; Grimmelman, 2015; Polonioli et al., 2023; Rahman et al., 2023), the construction of homes for the express purpose of experimentation and data collection (Taylor, 2021; Mollen, 2023), field tests with geoengineering technologies (Stilgoe, 2016) and open innovation practices that benefit from user innovators (Bogers et al., 2018).

These activities have practical and epistemic benefits, but they have also raised significant ethical concerns. For example, it has been pointed out that they can expose persons to risks or harm (Stilgoe, 2020), cause undue influence and manipulation (McDermott & Hatemi, 2020), violate their privacy or even human rights (Amnesty, 2020), or enroll persons in research against their knowledge or consent (Kitchin, 2016). These ethical concerns need to be investigated, and they have drawn much-needed attention to the ethical risks of the research formats mentioned.

However, ethical reflection on these various research activities has often been piecemeal and focused on domain-specific issues. Therein lies a problem. On the one hand, we might be overlooking a deeper ethical concern that unifies these seemingly varied and diverse new forms of research, and our moral and regulatory response may fall short. On the other hand, attention to domain-specific issues (e.g., surveillance, privacy, and bias in the case of predictive policing (Galič., 2019; Susser, 2021)) may come at the expense of finding guardrails that should be applied to all such research formats, notwithstanding their domain.

Therefore, this paper addresses the question of whether there is a shared, unifying, and ethically significant feature common to several similar, though not obviously related, forms of real-world research. We argue that this is the case. Several seemingly diverse forms of real-world research, ranging from self-driving car tests to online A/B tests, which have received independent but 'localized' (to wit, domain-specific) ethical scrutiny, are unified by a shared common morally salient feature, which we call 'coupling.' By identifying this morally salient factor common to all real-world

research formats, we pave the way for a comprehensive approach to the ethics of research conducted under real-world conditions.

We define coupling as when two potentially independent[13] options are 'coupled,' meaning you cannot choose (or reject) one without the other. Coupling occurs, for example, when research is conducted under real-world conditions in a public street, and a person can no longer engage with this environment – walk that street - without being subjected to the research[14]. Before the research was set up, people had the option of 'going to the street' without the option of 'being a subject.' After the experiment is set up, both options are coupled, and one can no longer accept or reject one without the other.

This chapter proceeds as follows. First, we illustrate several paradigmatic examples of the phenomenon we are discussing. After, we briefly consider and reject four other potential 'sources' for a shared salient moral feature: (1) due to their 'real-world '-ness, (2) lack of researcher control, (3) their status as research or experiments, and (4) absence of (informed) consent. We then present our core argument: that real-world research is united in coupling options available to persons and that this coupling is morally salient. A full moral account of coupling is outside the scope of this paper. However, we point out grounds for moral concern. Finally, we outline various promising avenues for future research.

2.2. Illustrating the phenomenon

In this section, we introduce several paradigmatic examples of various research formats conducted under real-world conditions. The aim is to fix ideas about the phenomenon we are discussing. In particular, we suggest that the following examples are all recognizable as forms of 'research conducted under real-world conditions.' This sets the stage for our project, namely, to find a unifying and morally salient feature. Consider the following examples:

---

[13] With independent, we mean conceptually and metaphysically independent, i.e., being a bachelor and an unmarried man are not potentially independent (conceptual), and being h20 and water are not potentially independent (metaphysical).

[14] Throughout this chapter, we use (research) subjects rather than (research) participants. This terminology does not assume that they voluntarily engaged, are informed, etc. They are subject solely in virtue of being *subject to* the research, either being causally influenced by it or being observed in some sense, e.g. through data gathering.

CARS: In 2015, Uber started testing autonomous vehicles on public roads in select US cities like Pittsburgh and San Francisco. As Jack Stilgoe noted: "This was artificial intelligence *in the wild*: not playing chess or translating text but steering two tonnes of metal" (2020, p. 2, italics our own). In Arizona, Uber enjoyed little regulatory oversight and tested the cars without informing the public (Stilgoe, 2020). After a fatal accident in 2018, in which a pedestrian was struck and killed, Uber suspended its self-driving tests in Arizona (Stilgoe, 2020).

VIOLENCE: Stratumseind is a popular street and nightlife center in the Dutch city of Eindhoven and a hotbed for violent nightlife-related incidents (Mollen, 2018). Aiming to reduce this violence, Dutch Police conducted predictive policing experiments in Stratumseind. Predictive policing refers to a collection of (digital) technologies that use data and algorithms to predict and prevent crime from happening - or at least cut down response time significantly (Susser, 2021). For example, video and sound data were analyzed for signs of aggression, and social media activity in the area was analyzed to determine visitors' moods. Police officers were alerted in real-time on their smartphones about locations with potential for violence – allowing them to intervene sooner.

BURGLARY: The Fieldlab Burgarly Free Neighbourhood was a project in the Dutch city of Rotterdam focused on detecting suspicious and potentially criminal behavior and subsequently influencing this behavior (Ragazzi et al., 2021). One case within the project focused on analyzing video imagery to estimate the presence, position, and direction of people in order to detect anomalous and suspicious trajectories that deviate from the gathered dataset (Ragazzi et al., 2021). Another case focused on the detection of suspicious sounds – for example, the sound of breaking glass – and sending warnings through speakers or an automatic signal to nearby residents about this 'suspicious' behavior (Ragazzi et al., 2021).

SHOPPING: In 2024, Dutch supermarket conglomerate Jumbo announced they would start with several tests using artificial intelligence to identify and prevent shoplifting (van Monsjou, 2024). The software analyses camera footage for suspicious or divergent behavior and subsequently alerts personnel. They intended to roll out the technology nationwide if the tests proved successful.

EMOTION: In 2012, a collaborative study was conducted between Facebook and Cornell University to research the spread of emotion through social networks (Kramer, 2012). Researchers at Facebook would alter the amount of positive or negative posts on the news feeds of certain users in order to see whether it affected the emotion of the user's subsequent post positively or negatively (Flick, 2016). The study generated substantial controversy. Facebook users were unaware that their feed was manipulated and did not consent to their participation. The Cornell University researchers had sought IRB approval, but since data collection was done independently before their involvement (by Facebook), the IRB judged that no review was necessary (Flick, 2016).

LOVE: In 2014, the dating website OkCupid conducted various 'mismatching' tests on their matching algorithm (Benbunan-Fich, 2017). This algorithm generated a certain compatibility percentage between people to indicate whether they are a 'good' or 'bad' match. In the experiment, OKCupid would take a pair of users deemed a poor match and alter their compatibility percentage to suggest they were a good match. Unsurprisingly, subjects would send more messages to these matches and were more prone to engage in conversation. After, affected users were notified about the experiment and the actual match percentage.

GEO: Geo-engineering, alternatively known as climate engineering, refers to the idea of deliberately intervening in the Earth's climate system to mitigate or offset the impacts of global warming and climate change (Stilgoe, 2016). Geo-engineering involves techniques such as carbon dioxide removal (CDR), which aims to capture and store carbon dioxide from the atmosphere, and solar radiation management (SRM), which aims to reflect a portion or reduce the amount of incoming solar radiation. SRM techniques vary from low-tech solutions, such as increasing the amount of reflective surfaces, to injecting certain aerosols into the atmosphere. In September 2011, a group of British scientists planned a field test that aimed to do the latter: deliver a few dozen liters of water a kilometer into the sky by attaching a hose to a helium balloon, where the water evaporates into a mist. The experiment, however, soon raised controversy and was called off (Stilgoe, 2016).

HOMES: The 'DreamHus' are so-called live-in laboratories - inhabited living environments constructed for research purposes (Mollen, 2023) - located at the

> Delft University of Technology campus (Green Village, 2021). Their inhabitants are both residents and research subjects. The houses aim to test potential innovative solutions to develop sustainable housing and subsequently scale up these solutions to the general (Dutch) housing stock. Tests done by researchers, students, and innovators within the homes include "solutions in the field of energy, healthy indoor climate, water, heating, insulation, ICT, IoT and Smart homes" (Green Village, 2021).

There are clear differences between these various research formats conducted under real-world conditions. They have different aims, methodologies, agents conducting them, the domain about which knowledge is sought, and the type of person affected, to name a few.

Similarly, ethical discussion about such interventions has often been piecemeal and focused on such domain-specific issues. For example, the case of self-driving cars has been amply discussed (Santoni de Sio & Van den Hoven 2018), as has predictive policing (Susser, 2021) or the case of geo-engineering (Stilgoe, 2016). Focus has often been on domain-specific issues, such as assigning responsibility and meaningful human control in the case of self-driving cars (Santoni de Sio & Van den Hoven 2018), privacy and bias in the case of predictive policing (Susser, 2021) or irreversible risks in the case of geoengineering (Pamplany, Gordijn and Brereton, 2020).

On the other hand, it seems clear that these phenomena are also related to one another in an important sense. So far, however, it has not been articulated with sufficient precision how, if at all, these phenomena are related. More specifically, it is unclear if there is any morally salient feature that all of these phenomena share. At the same time, it is of ethical significance to uncover such a feature because it might ground a unified ethical response. In other words, their relatedness might justify a unified ethical approach, and understanding it better might reveal important ethical insights. In the next section, we consider some initial ideas about this unifying factor.

2.3. Rejecting four potential unifying moral features

In this section, we consider and reject four potential unifying and morally salient features of the above phenomena: (1) their 'real-world '-ness, (2) lack of researcher control, (3) their status as research or experiments, and (4) absence of (informed)

consent.[15] In what follows, we show that each proposed feature in this section fails to descriptively capture the phenomena we are interested in or, if it does, it is of unclear moral relevance and thus not fruitful to underwrite an overarching moral inquiry into real-world research.

2.3.1. Real-world conditions

One common factor that all the above-mentioned cases share is that they are conducted, in some sense, under real-world conditions. These conditions are also referred to as 'everyday social contexts,' 'natural environments,' 'in the wild,' or as research that is conducted in 'real-world' or 'real-life' contexts (Kroes, 2016; Ansell and Bartenberger 2016, 2017; David & Gross, 2019). One reason that favors 'real-world conditions' as a unifying factor is that their being 'in the wild' is their reason for existence. Often, they are conducted because the 'uncontrolled' complexity of the conditions is necessary to acquire relevant knowledge of how the intervention will operate within their eventual potential use-setting.

However, 'real-worldness' does not seem to be a morally salient feature. One reason for this is that there are different kinds of 'real-world' conditions or 'use settings' under which real-world research can operate. These contextual differences can be morally salient in and of themselves and, therefore, unable to ground a moral concern shared by all of these examples of real-world research. For example, while CARS, VIOLENCE, and BURGARLY operate in public *physical* spaces, EMOTION and LOVE operate in a virtual and private environment. Additionally, while CARS, VIOLENCE, and BURGLARY operate in public spaces, SHOPPING, EMOTION, and LOVE are conducted in commercially private environments. Furthermore, whereas CARS, CRIME, EMOTION, and GEO intervene in previously existing environments, the 'real-world' conditions of HOME

[15] We want to find a common feature of real-world research that satisfies two desiderata. First, the feature is shared by all of the observed phenomena. Second, the feature is morally salient in the right way. By that, we mean that the feature conceivably gives rise to ethical concerns that are sufficiently closely tied to the unifying factor without being too general or broad. To illustrate, a feature that does not ground justified ethical concerns is not morally salient. By contrast, a feature that does ground justified ethical concerns but is based on general features that are not at all unique to the kind of phenomena we are investigating here also does not count as morally salient in the right way.

were created purely for research purposes and only 'made' into an everyday social context after the fact.

What is more, focusing only on the different ways these experiments are 'real-world' makes it doubtful that all of them are morally problematic. To illustrate, though it would seem problematic to use data won in a public setting, as in VIOLENCE, for commercial purposes, the same is not necessarily true for data won in a private, commercial setting, as in EMOTION. Thus, insofar as there is a moral concern at all, it does not arise from the fact that they are conducted under real-world conditions.

2.3.2. Lack of researcher control

In contrast to controlled laboratory research, all the above cases are characterized by (varying) lower degrees (or even an absence) of controls by the researcher on the environmental research conditions (Harrisson, 2005; Ansell & Bartenberger, 2017; Kroes, 2016). Here, the notion of control refers to the control investigators have over the experimental intervention and environment to preserve the epistemic validity of a particular research outcome, such as randomization, control groups, isolating variables, or specific statistical methodologies (Ansell & Bartenberger, 2017). This lack of epistemic control is interesting since the production of knowledge is one of the main reasons for conducting real-world research in the first place.

While we are inclined to agree that lack of researcher control unifies those cases, with the two caveats that this can change between research environments and that the boundary or threshold is unclear for what constitutes 'low control,' it should be clear that it is not an ethically salient criterion by itself. A lack of epistemic control is too ubiquitous. The vast majority of activities in life are conducted under real-world conditions or without epistemic controls: taking the train, baking, writing, sailing, skydiving, etc. However, these activities do not seem morally salient due to the fact that they are conducted under real-world conditions or in the absence of a particular epistemic control mechanism over the researcher environment by researchers. Moreover, we can see that this is not a morally salient feature because we can imagine a scenario where, for example, self-driving cars are tested in a real-world environment that is perfectly controlled by the researcher through some mind control drug he uses with all subjects. That would, if anything, only worsen the moral issue.

2.3.3. Research and Experimentation

All of the examples are (or at least have been called) forms of research or experimentation. However, this is not a helpful unifying feature either. First, there exists conceptual ambiguity and disagreement about what constitutes research or experimentation. For example, the concept of experimentation is applied to various practices inside and outside scientific inquiry.[16] Consequently, depending on which notion we use, some cases are included, but not all. At the same time, an all-encompassing notion – for example, that these are all activities with the express purpose of learning something – would be too vague.

Even if we identify or settle on a clear criterion of 'experiment' and find that all these phenomena are experiments, that in itself is not ethically interesting**.** Of course, experimental real-world research can raise similar ethical concerns as all other forms of experimental research. However, such issues would be morally salient ways in which real-world research can occur and are not necessarily constitutive of real-world research in and of itself. More importantly, the cases we discussed evoke the idea that there is something special concerning the ethics of real-world cases, over and above the ethics of research and experimentation per se. This would be lost if we just focus on their research or experimental nature.

---

[16] Ansell and Bartenberger distinguish between 'restrictive' and 'expansive' notions of experimentation (2016). More 'restrictive' conceptualizations of experimentation mainly refer to experimentation as a (scientific) research method that equates experimentation with 'control techniques' such as randomization, control groups, and isolating variables (Ansell & Bartenberger, 2016). 'Expansive' positions towards experimentation are broadly more concerned with innovation and realizing certain desirable states of affairs and, in some cases, are defined by their *absence* of techniques of control. Under this expansive position, for example, experimentation is used to describe processes of iterative problem-solving and approaches to improving policy, governance, institutions, and technology (Ansell & Bartenberger, 2016). Several authors have criticized broader notions of experimentation for losing analytical value (Ansell, 2019, p. 11; Gross, 2010, p. 66; Karvonen & Van Heur, 2014). Other times, experimentation is used in a more metaphorical sense. For example, Hansson notes that: "some academic writers, typically among those not attending much to methodological niceties, also use "experiment" in a wide sense that does not imply planned interventions and observations, but to the contrary puts focus on the lack of planning and monitoring in certain human activities. This applies, for instance, to descriptions of potentially harmful practices and developments" (Hansson 2019, notes 1).

2.3.4. Lack of consent

Having rejected some preliminarily plausible criteria, we suggest that a more promising criterion is that they all research phenomena that *lack* informed consent: that a competent individual has received information about the research, has understood this information, and has been able to decide on participation free from undue influences (Rhodes, 2005; Hansson, 2006). This is a plausible proposal since the above phenomena seem to involve persons in research indiscriminately, regardless of whether they are aware of or agree with such inclusion. Additionally, when conducting research on a group level, subjects are not recruited individually, and therefore, there might not be an immediate opportunity to inform or obtain consent from individuals. Alternatively, as various authors have pointed out, there might be significant challenges to identifying all those who might be involved or affected by the intervention in a real-world context (Schinzinger & Martin, 1983; Hansson, 2006; Van de Poel, 2016)[17].

While we do agree that the presence or absence of consent in research is morally salient, we do not think that a lack of consent is what unifies the cases mentioned above. First of all, while consent is absent in some cases, other cases *do* involve consent – in some cases more explicit than others. The residents in the case of HOME actively sign up to live in their experimental houses. The users of EMOTION and LOVE might have signed off on being part of an experiment within the terms of service, even if that consent might be hidden deep within the terms of use (Pfotenhauer et al., 2022). It thus fails the representative condition.

However, even if we would remove these exemplary phenomena and only look at cases in which consent is truly absent, the absence of consent remains not what is uniquely morally interesting about these cases. Any research format can be imposed on persons without their consent. For example, hospital patients can be part of a highly controlled experimental drug trial without their knowledge or consent. More importantly, if we were to imagine that consent is obtained in all the above cases, it would mean that it would void any morally interesting feature unifying the cases. We do not think this is the case. We argue that *independently* of the presence of informed consent, there remains a feature that remains descriptive and morally salient. We expand on this in the next section.

---

[17] Furthermore, depending on the nature of the study, researchers might fear that obtaining informed consent might undermine the 'real-world' conditions of the experiments and thus negatively affect the epistemic quality of the study and thus forgo obtaining consent from people.

2.4. Coupling Options

In this section, we consider and reject four potential unifying and morally salient features of the above phenomena: (1) their 'real-world '-ness, (2) lack of researcher control, (3) their status as research or experiments, and (4) absence of (informed) consent.[18] In what follows, we show that each proposed feature in this section fails to descriptively capture the phenomena we are interested in or, if it does, it is of unclear moral relevance and thus not fruitful to underwrite an overarching moral inquiry into real-world research.

So far, we have reviewed and rejected four potential features that are both representative of all real-world research cases and are (potentially) morally salient. In this section, we defend an alternative view. We argue that real-world research is unified in a feature that is morally salient in the right way: option coupling. In a nutshell, our thesis is that real-world research *couples* options available to subjects. With options, we are concerned with options for actions. We define coupling as when two potentially independent options become coupled so that you cannot choose (or reject) one without the other.

We will now explain our account in more detail. In all the above-mentioned cases, even if they are aware or knowingly consent to be part of one of the research projects (e.g., a sign on the street notifying residents and passersby of the research), people are faced with the choice: take the option of engaging with the particular research environment (walk the street, use the website, etc.) and be a research subject, or the option of not being a research subject, but then also forgo engaging with the respective environment completely.

Scholars discussing aspects of real-world research have made reference to this dynamic in passing, although fragmented and not in great depth. For example, in the context of smart city research, Kitchin writes that people often have "*little choice* in being surveilled … *there is no ability to opt-out other than to avoid the area*, which is unreasonable and unrealistic. As such, there is no sense in which a person can

[18] We want to find a common feature of real-world research that satisfies two desiderata. First, the feature is shared by all of the observed phenomena. Second, the feature is morally salient in the right way. By that, we mean that the feature conceivably gives rise to ethical concerns that are sufficiently closely tied to the unifying factor without being too general or broad. To illustrate, a feature that does not ground justified ethical concerns is not morally salient. By contrast, a feature that does ground justified ethical concerns but is based on general features that are not at all unique to the kind of phenomena we are investigating here also does not count as morally salient in the right way.

selectively reveal themselves; instead, *they must always reveal themselves*." (2016, 9, italics our own). In the same context, Zimmermann writes about the challenge of obtaining consent in real-world research practices and notes that, for example, a "privacy notice attached to the streetlamp might prove insufficient to inform all affected citizens about the data collection and *give them the opportunity to opt-out by avoiding a certain area*" (2023, p. 53; italics our own). These are valuable starting points from which we develop a fuller view of the phenomenon of coupling now.

In working toward a fuller understanding of coupling, we bracket the normative question about when or whether such choices are 'unreasonable' or 'unrealistic.' Instead, we aim to build a descriptive explanation of the dynamic at play. First, rather than saying that people only 'must always reveal themselves' (per Kitchin), it would be more accurate to say that it is impossible to *selectively* reveal oneself *within* the scope of real-world research. There, one has to accept or reject the situation wholesale. Second, it is clear that people can still choose to reveal themselves selectively by avoiding the locations of data capture (rejecting the option of engaging with the respective environment, in our terminology). Therefore, the cases do not seem to affect a person's choice-*making* capacities. Rather, it changes the options available to them in some way.

To illustrate what it means that options are coupled, consider again the case of VIOLENCE, as mentioned in Section 2. A popular nightlife street has been transformed into a testing site to test various emerging technologies. One such technology is lighting panels that aim to lower aggression through colored light. After an initial laboratory study, the lighting panels were tested in a real-world environment (the nightlife street) for a period of two years (Sens, 2015)

Before the experiment began, a person's relevant options (heavily abstracted) might look like this:

Option A: Go to the nightlife street.
Option B: Do not go to the nightlife street.

Similarly, a person's options regarding research participation in an experimental lighting test within a controlled environment, as in a university's research lab, might have looked like this:

Option C: Be part of a test on experimental lighting
Option D: Don't be part of a test on experimental lighting.

Importantly, the options A, B, and C, D were (and have the potential to be) independent of one another. The effect of the experiment outside the lab is that this is no longer the case. Instead, in the above example, a person is now presented with the following *coupled* options:[19]

> Option A+C: Go to the nightlife street AND be subject to the lighting test
> Option B+D: Do not go to the nightlife street AND not be part of the lighting test

Clearly, real-world research 'couples' the options available to individuals[20]. Picking option A now also necessarily involves C. Alternatively, picking A excludes D and vice versa. If someone does not want to be involved with the lighting experiment (Option D), they must avoid the nightlife street (Option A). Picking both Option A (Go to the nightlife street) and Option D (Not be part of the test) has become impossible.

By coupling options, one interferes with persons in a particular way (even if it does not affect them in a material sense). If these options used to be independent, coupling real-world research deprives people of some options. Whereas once a person has the option to engage with a particular environment or activity without being part of a research project, this particular option is no longer available to them.

When the coupling is not transparent, one deprives persons of a 'genuine' choice (a choice in which all the relevant information is available to them) to weigh the benefits and risks and decide on participation by avoiding the research location. Consequently, people who are not properly informed about the coupling can become unwitting subjects. For example, in the EMOTION and LOVE cases, users were (initially) unaware that their feed was manipulated for research purposes

---

19 We refer to the relationship between scenarios such as A+C and B+D as *coupled* options, in that you cannot pick one without the other. You cannot pick option A without also 'getting' option C (or, inversely, reject C without also rejecting A). We will refer to mechanisms that 'decouple' coupled options as *decoupling*: to separate coupled options into two or more distinct options that can be chosen independently.

20 One question about our terminology is whether options need to be conscious, i.e., as something that the agent is aware of and can choose. In that sense, if you do not know that A includes B, you did not really *choose* A+B. What matters to our account is the conditions that an option subjects a person to and not whether they consciously chose this option. Whether or not they are aware, it would not alter the fact these options are coupled together.

(Kramer, 2012). Similarly, in CARS, residents of Tempe were not notified that the experiment was taking place within their city (Stilgoe, 2020).

Additionally, despite the subject's awareness or consent, by coupling options researchers have altered the values associated with the coupled options. This can be thought of as if the researchers conducting the experiment have added a particular (transparent or untransparent) *cost* to the options available to persons. That is, refusal to participate comes at the loss of being able to visit a particular environment (e.g., avoid a nightlife street). This loss a person would not have to incur were the options decoupled. Furthermore, this cost of refusing to participate in a coupled research format might influence a person to not reject participating in the research since persons might miss out on something they previously had access to.[21]

In summary, coupling is the phenomenon of two potentially independent options becoming coupled so that a person cannot choose (or reject) one without the other. We consider our discussion of coupling to represent initial steps toward a fuller appreciation of the phenomenon in general and, as we argue later, of seeing its relevance for understanding the ethics of real-world experimentation. There are important limitations to our discussion. Most importantly, we represent coupling in fairly general terms focused on the *options* available to people. While we take that to be an intuitive representation that is suitable for present purposes, a fuller treatment will have to make these notions more precise, for which one could draw on suitably developed frameworks, e.g., in decision theory. For now, we will turn to show that our broad construal of coupling already serves to uncover unity amongst the forms of real-world experimentation discussed in Section 2.

2.4.1. Representative Condition

Coupling is a unifying, identifiable feature since similar dynamics arise in every example of real-world research that we discussed at the beginning of this paper. This is true whether the research is conducted online (not being part of an A/B test means avoiding using particular social media), on public roads (not being part of a self-driving car experiment means avoiding public roads where it drives), or on private property (to not to be part of an AI-assisted shoplifting prevention test, you have to shop elsewhere). To illustrate this, we will briefly discuss how each case

[21] This does not mean that by participating in research, subjects might not gain particular benefits.

mentioned in Section 2 (aside from VIOLENCE, which we already discussed) involves coupling.

CARS: Testing autonomous vehicles on public roads involves coupling in that when the persons engage with traffic, they become research subjects to the development of the experimental vehicle. Not being part of a self-driving car experiment means avoiding public roads where it drives. The choice to partake in traffic is coupled with being exposed to the test and its risks.

BURGLARY: Testing technologies for detecting and influencing potential criminal behavior within a residential neighborhood means that residents, visitors, or passersby cannot visit the parts of the neighborhood where the technology is located without being a subject in this research since these options are now coupled together.

SHOPPING: Shopping in a supermarket that runs tests with AI-assisted shoplifting prevention means you are subjected to that test. Due to coupling, one cannot shop and not partake. If one does not wish to participate, you must shop elsewhere.

EMOTION & LOVE: Engaging with an online platform such as Facebook or OKCupid that runs tests on their users is impossible without being subject to research. Not participating means avoiding the online platform since these options are coupled.

GEO: Intervening in the Earth's climate system to mitigate or offset the impacts of global warming and climate change means that all those who live within that climate system are subject to the research. Since these options are coupled, in order to avoid the research intervention and its effects, one would have to avoid our climate system altogether.

HOMES: It is impossible to live in a home that is tested upon and not also be subject to the research. These options are coupled, and to not be a research subject, one has to move.

One objection to our argument is that this coupling is hardly unique to the examples of real-world research we provided but is, in fact, a ubiquitous part of life.

Take, for example, toll roads. In order to access a toll road, one has to pay a fee or toll. If one refuses to pay this toll, they cannot use the road (and must use other ways to reach their destination). Regardless of specific circumstances, one must pay a fee to enter, and one cannot separate the choice to engage with the road (A) without the choice to pay (B). As the toll road exemplifies, this dynamic also exists in many other practices.

However, what differentiates the problematic type of coupling in real-world research from coupling in, for example, a toll road are further features. One clear feature is transparency. In the toll road example, riders are informed about the coupled options. This is not the case in several of the examples of real-world research that we discussed. Another distinguishing feature concerns the independence of the choices. Insofar as a toll road is built at all, there has never been a de-coupled choice of taking the road but not paying. There is no choice that is taken away from a person. This is another differentiating feature.

Naturally, we can conceive of coupled choices that are not exclusive to real-world research. Our criterion might thus overgeneralize. But we do not want to be too quick with that conclusion. Insofar as a road becomes a toll road, we do face a coupled choice. The effect will likely be similar to real-world research. And we might be justified in asking the same kind of ethical questions that we ask about real-world research. Therefore, we suggest that coupling is a useful criterion that is a) shared by all cases of real-world research that we discussed (and which, we argued, are representative of a class of phenomena that are interesting from an ethical perspective for similar reasons) and b) it might be pointing to further phenomena of moral salience that are not readily conceived of as real-world research but which raise similar ethical questions.

2.4.2. Moral Salience

What is the issue with coupling options in this way? A detailed moral account of coupling is outside the scope of this paper. However, we argue that the moral salience of coupling is, at least, conditionally dependent on the moral salience of (and the interplay between) the options themselves that are coupled, as well as the degree of control one has over the coupling in the first place. In other words, the 'moral weight' of the coupled options that one has to accept or reject gives moral salience to the dynamic of coupling. For clarification purposes, in this section, we distinguish

between the 'live' option (previously option A) and the option that gets coupled to the live option as the 'experimental' option (previously option B).

First, there is the live option and its importance to the subject. One of the characteristics of coupling is that if one wants to reject the experimental option, one also has to forgo the live option. Depending on its content, this adds moral salience to forgoing the live option. For example, it seems more morally salient when the live option is the option of being able to participate in public life (traffic or public space like in CAR or VIOLENCE) rather than whether it represents a more trivial option, such as being able to go to a specific shop that is very common (such in SHOP). When the live option is more important to a person, yet I have to accept additional conditions (the coupled option) to receive access to it, the coupling seems more morally salient.

Thus, the moral significance of the live option, which can express itself in a variety of ways, will influence the significance of coupling such that *the more significant* the live option is, the more we should be suspicious of coupling it with a coupled option, i.e., some experiment. Although this broad trend could be specified much more, it already provides a helpful lens through which we can assess the morality of different forms of real-world research. For example, the degree to which it is problematic to couple social media use with experiments seems to depend in part on the costs of avoiding social media use (of which there is a rich descriptive and normative discussion). Similarly, our coupling perspective suggests that we should view with much greater suspicion experiments that couple weighty, significant choices like where one lives with an experiment, as opposed to fairly insignificant choices such as which street one visits in a city. Naturally, if visiting that street holds great significance for the individual, e.g., because it is where she lives, works, or needs to cast a vote, the coupling becomes much more pressing, in line with our analysis.

Second, the moral salience of coupling seems to also depend on the nature of the experimental option. It seems more morally salient when a person visits a particular environment that they now have to accept conditions that are very risky, harmful, or not beneficial to the public interest compared to whether the coupled option involves a (morally) uninteresting observational study that involves lower degrees of risk, harm, or disadvantages to the public. We can thus formulate another prediction based on our coupling perspective. All else being equal, the more negative moral salience that comes with the experimental option, the more we should view the resulting real-world research with suspicion. Again, this is a highly general formulation, but our point is that it already helps to recognize shared moral issues with the

different types of real-world research that we discussed in section 2 (and, plausibly, more that we did not explicitly discuss in this paper). All else being equal, if we introduce risks to people's lives, as in, for example, the public testing of self-driving cars that are far from technically mature, by way of a real-world experiment, that is ethically more troubling than introducing the comparatively lower harm of a minor privacy intrusion in case of an observational study 'in the wild.' Importantly, the ethical assessment of both phenomena has a shared ground, namely the nature of the coupled option.

Another concern involves what personal control subjects have *over* the coupling in the first place. Personal control is broadly defined here, and we use it to capture concerns that relate to the knowledge, control, and consent a subject has over whether particular options are coupled (or decoupled). For example, have people been informed that a particular environment is now a research environment? Have they been consulted about this, or has their consent been obtained? Do they have the ability to 'decouple' these options and continue engaging with the (now) research environment without being a subject? It seems that coupling becomes more morally salient insofar as one lacks control. In that case, one cannot protest the coupling nor work for decoupling.

The interplay between these conditional morally salient features seems to greatly affect the overall moral salience of coupling. For example, whether all, some, or none of the options are morally salient clearly affects the degree of moral salience of the coupling at hand. Take, for example, the worst-case scenario. Say, the live option is of great importance to me, *and* the experimental option is tremendously severe, *and* I had no knowledge or control over the coupling in the first place nor have the control to decouple the options; such a case seems *more* morally salient than, for example, when the live option is of no great importance me, and the experimental option is not particularly severe, *and* I gave my consent about the coupling and are empowered to decouple them.

At this point, one might question what is morally salient about coupling itself if its moral salience depends mainly on the moral saliency of its parts. However, it is not *merely* the moral salience of options. For example, being a subject of a risky experiment that influences your emotions in a laboratory setting is one thing. Being subject to the same experiment but which is coupled with visiting a particular nightlife street is another. With the first, one never had access to this laboratory environment to begin with, as they might have had with a public street or square that is now altered for research purposes, and thus, they do not need to accept additional conditions or

forgo this access to avoid the research conditions. Hence, the moral salience coupling rests both in the moral salience of the options involved *and* the fact that they are coupled.

Despite our focus here on conditional moral salience, we do not want to discount the possibility that coupling has substantial normative significance, i.e., moral significance above and beyond all the conditional moral salience of coupling. Coupling does seem to present a particular constraint on subjects, yet whether this constraint is morally salient in even seemingly trivial cases is a question outside the scope of this paper. However, in the next section, we outline some avenues for further research that might shed further light on this issue.

2.5. Discussion

Our account of coupling presents a novel and ethically fruitful analysis and paves the way for a comprehensive approach to the ethics of real-world research. However, the scope of this paper is limited, and further work is needed to build a fuller philosophical account of coupling. In this section, we outline some interesting avenues.

First, in the previous section, we mentioned that coupling seems particularly morally salient when one has no control over the coupling. A promising lens through which to approach this question is the neo-republican accounts of freedom. Neo-republican thought stresses "people's 'effective control' over their life as an independent dimension of evaluation of well-being" (Santoni de Sio, Txai Almeida, and Van Den Hoven 2024, p. 670). In contrast with liberal traditions that define freedom as freedom from *actual* interference (Berlin, 1969), neo-republicanism defines freedom as being free from (or the absence of) the *potential* interference of arbitrary (or *dominating*) exercises of power (Lovett, 2010). Briefly put, non-arbitrary power is controlled by either effective rules such as the law (procedural position) or concerned persons and groups themselves (democratic position) (Lovett, 2022). Recently, Capasso has used this theory to argue that digital nudges— features of user-interface design, like dark patterns, that guide people's behavior in online choice environments (Weinmann, Schneider, and Brocke 2016)— are wrong when and because they *dominate*: they actualize an influence that is not *controlled* by the people that they target (2022). Similarly, Maheshwari and Nyholm argue that certain risk bearers, for example, those exposed to experimental self-driving vehicles in their neighborhoods (like CARS), suffer from what they call 'dominating risk impositions' which constitute "a problematic form of relationship between those who have uncontrolled or

unchecked power to impose risk and those who are vulnerable to its imposition" (2022 p. 615). Such accounts offer a promising starting point for further analysis.

A second avenue concerns what considerations could override morally problematic coupling. Within the research ethics literature, it is generally taken that people should not be forced or coerced to be part of the research[22] and that research participation should be conceived of as a free and voluntary activity (London, 2020). However, such considerations are not taken to be absolute. For example, people often have little rights over their data being captured in a public setting (Spicker, 2011), and in certain cases, research without informed consent is considered ethical (Gelinas, Wertheimer, and Miller, 2016). People have seemingly limited rights to control their participation or data when both are the 'by-product' of their natural behavior and when data are anonymous. Coupling, while morally salient, might, therefore, not always cause real-world research to be unethical. Hence, clear conditions must be articulated for when unethical coupling can be overridden by other considerations.

A third line of questioning concerns how different real-world research environments offer seemingly different opportunities to decouple options. For example, online real-world environments are seemingly better tailored to offer specific avenues for personal control than 'physical' real-world environments. The same website can be shown with slight alterations to various persons, each simultaneously interacting with a distinct version of the same website, yet this is clearly not possible for a physical location like a public square. What is seemingly possible is that in some cases, specific individual interactions (for example, interacting with a specific biometric border gate in an airport (Schiphol, 2017)) can be 'decoupled' from other interactions in the same physical space.

2.6. Conclusion

In this paper, we have argued that seemingly separate phenomena of real-world research – like tests with self-driving cars in inner cities, smart city interventions, and online A/B tests – are unified by a common and morally salient feature, which has received little scholarly attention so far: real-world research 'couples' options. By

[22] For example, the Nuremberg Code states that research subjects "should be so situated as to be able to exercise *free power of choice*, without the intervention of any element of force, fraud, deceit, duress, over-reaching, or other ulterior form of constraint or coercion" (Shuster 1997, p. 1436).

testing interventions under natural conditions in a particular environment, an (involuntary) 'subject' can no longer engage with that environment without being part of the research and the associated risks, influences, and data capture. They can no longer pick option A (the live option) without accepting option B (the experimental option). Alternatively, they have to reject both. The 'moral weight' of the coupled options that one has to accept or reject gives moral salience to the dynamic of coupling. While our scope was limited, and more work is needed to build a fuller philosophical account of coupling, our account presents a novel and ethically fruitful analysis, paving the way for a unified approach to the ethics of real-world research.

Prologue to Chapter 3

The previous chapter offered a comprehensive account of the ethics of real-world research. It identified various unifying factors of real-world research and argued that 'coupling' is a unifying and morally salient factor. In doing so, it laid the groundwork for a unified approach to the ethics of real-world research.

In Chapter 3[23], I discuss the general lack of research ethics governance in real-world research. With this, I mean that there are few governance structures in place to regulate relevant examples of real-world research and ensure they are conducted in accordance with particular regulatory or ethics standards. Comparatively, we place such regulatory demands on scientific research, meaning we hold different research types to different regulatory standards. In the next Chapter, I focus on research ethics governance and ask whether it is justified that we hold real-world research to fewer (or none) research ethics standards compared to much of scientific and publicly-funded research. Specifically, in Chapter 3, I aim to answer the following question.

> RQ3: *What, if anything, justifies the lack of research ethics governance for real-world research in comparison to scientific research?*

I will argue that the current situation is unjustified and that we have good reasons for remedying inconsistent research ethics governance demands between research domains. However, in this Chapter, I remain descriptive regarding whether existing research ethics principles or norms *should* apply to real-world research as a matter of morality. Instead, I show that, irrespectively, scholars have started to use these research ethics principles and norms to evaluate real-world research and, thus, note that there are, at the very least, beliefs about these research ethics norms applying (which might be true or false). Rather, it aims to argue that real-world research brings about apparent ethical challenges that a research ethics governance should address and mitigate and that a current absence of ethical governance is, therefore, problematic.

---

[23] Published as: Mollen, J. (2024). Towards a research ethics of real-world experimentation with emerging technology. *Journal of Responsible Technology*, 20, 100098.

3. Towards a Research Ethics and Governance of Real-World Research

Abstract

Testing emerging technologies, such as autonomous vehicles, predictive crime analytics, and smart city interventions under real-world conditions is an important strategy for robust and responsible technology development. However, the research ethics of real-world research often remains unaddressed and unregulated. This article argues that there are problematic inconsistencies in the demands of research ethics governance across different categories of research and development with emerging digital technologies. This differential treatment is problematic since there are no meaningful differences to justify it, and it creates the possibility of regulatory evasion at the cost of populations' due protection. Hence, I argue that this differential treatment (i.e. lower research ethics regulatory demands) should be amended by developing research ethics governance for real-world research. In doing so, this paper contributes to several ongoing scholarly debates on the limits of current research ethics guidelines in the face of novel technologies and research formats.

3.1. Introduction

Testing emerging technologies under real-world conditions has emerged as a distinctive data-driven strategy to address various societal and innovation challenges in recent decades. This 'real-world' research and development is understood as an important strategy to bridge performance in controlled laboratory environments to eventual successful real-world deployment and to study to what degree a particular emerging technology can be leveraged to solve a particular social issue (Ansell & Bartenberger, 2016; Pfotenhauer et al., 2022). Examples range from online experiments such as algorithmic A/B tests with emotions, romantic relations, careers, and wages on social media or online worker platforms (Grimmelman, 2015; Wood, 2020; Polonioli et al., 2023; Rahman et al., 2023) to experimentation in (urban) physical environments, with self-driving cars on public roads (DeArman, 2019; Stilgoe, 2020), predictive policing and crowd control technologies in nightlife streets and neighborhoods (Galič, 2019; Amnesty, 2020; Susser, 2021), smart homes (Taylor, 2020; Mollen, 2023) and smart city interventions (Zimmermann, 2023).

While real-world research might benefit the development of responsible emerging technologies (Colonna, 2023) or help solve social challenges (Ansell & Bartenberger, 2016), attention should also be paid to conducting these experiments *ethically*. Since real-world experiments operate closely to people's daily lives or environment and actively intervene within them, they potentially cause undue influence, impact, or harm to persons who become (un)knowingly or (un)desirably involved. For example, Colonna notes that artificial intelligence "that is tested in real-world conditions" .. can present "risks to individual's health, safety and fundamental rights, as well as broader societal concerns" (Colonna, 2023, p.28). For example, testing experimental AI facial recognition systems that turn out to be biased can harm individuals or groups of people through discrimination, but also, as Smuha points out, can cause broader 'societal' harms such as "a higher interest to live in a society that does not discriminate against people based on their skin color and that treats its citizens equally" (Smuha, 2021, p.6).

However, while testing emerging technologies under real-world conditions raises various ethical issues, scholars have increasingly pointed out that there are only limited governance structures in place to help regulate these ethical concerns in real-world research. For example, in the context of online corporate experimentation, Poloni and colleagues observe that while "protection protocols have become the norm in medical research and the social and behavioral sciences," … "the use of human subjects in research that is not federally or publicly funded—such as in the

case of privately funded A/B testing, often affecting millions of potentially unaware people—has remained unregulated" (Polonioli et al., 2023, p. 669). As this paper will show, similar significant discrepancies exist between the regulatory demands placed upon different research and development categories, resulting in unequal treatment for researchers and research participants.

In this paper, I argue that these differences in research ethics governance are problematic because there are no meaningful conceptual differences between currently ethically regulated and unregulated research that would justify this difference. Additionally, due to the collaborative nature of modern research practices, this absence of research ethics governance for real-world research creates the possibility of research ethics regulatory evasion, which comes at the expense of those whom these demands aim to help protect. Consequently, I argue that these discrepancies should be amended by developing a research ethics governance for real-world research and harmonizing regulatory demands.

In doing so, this paper contributes to (1) the increased awareness about the research ethical dimensions of real-world research with emerging technologies (Taylor, 2020; Zimmerman, 2023; Polonioli et al., 2023; Rahman et al., 2023), (2) to various ongoing debates on the limits of existing research ethics guidelines for real-world research within AI and data science and social and political science, and (3) to ethics-by-design approaches that have argued that ethical considerations should be included in the design and development process of new technological by drawing attention to a category of research and development that need 'research' ethics-by-design (Dainow & De Brey, 2010; De Brey & Dainow, 2023).

This paper proceeds as follows. In the first section, I draw from two ongoing debates on research ethics reforms regarding real-world research within AI and data science and social and political science to ground my claim within a more extensive debate on the current shortcomings of research ethics guidelines. In the second section, I focus on the increasing attention to the lack of research ethics governance for real-world research. In the third section, I argue that the current situation is problematic since no meaningful reasons justify it. In the fourth section, I argue that this differential treatment creates the possibility of research ethics regulatory evasion due to the collaborative nature of modern research practices. Finally, I briefly discuss several avenues and challenges to resolving this issue.

3.2. The growing need for a real-world research ethics

In this section, I describe the moral salience of real-world research and discuss two ongoing debates on the need for research ethics reforms in field experimentation, specifically from the perspective of AI and data science and social and political experimentation. By connecting these two debates, I ground my later argument about harmonizing research ethics principles and practices with currently unregulated research in a broader ongoing academic debate about research ethics reforms.

As discussed in Chapter 2, real-world research, as the name implies, involves conducting research and testing under 'real-world conditions.' It differs from laboratory or controlled research since no experimental controls are placed on the research environment. We can distinguish between observational and experimental real-world research. The first studies phenomena that arise naturally; the latter actively brings about the phenomena studied through active intervention[24]. A theoretical example of the former would be using digital technology to capture location data to measure crowd density; an example of the latter would be studying how the use of various phrases on public digital billboards could influence crowd density. Both observational and experimental real-world research raise ethical concerns. In both cases, researchers place themselves or whatever they study in a participant's or community's daily lives or environment (Teele, 2014). However, real-world experimentation is responsible for creating the data they observe. Consequently, by intentionally altering the environment, they can bring about undesirable, unintended, and unforeseen consequences caused by the research intervention (Teele, 2014, p. 119).

Scholars increasingly call for research ethics reforms regarding the design or conduct of real-world research. One domain in which this debate is prominent is within AI and data science. Scholars have discussed the ethical and regulatory challenges for researchers and institutional review boards regarding social media and online data research (Moreno et al., 2013; Vitak et al., 2016; Raymond, 2019, p.

[24] This distinction between observational and interventional research is not necessarily sharp. One reason is conceptual unclarity about what qualifies as a manipulation or intervention in the research context. For example, while ethnographic research is considered observational research, the presence of researchers within these communities can influence the observed behavior. One way to resolve this, is to fall back on the researcher's intentions. As Teele puts it, this difference concerns the degree to which "the researcher *purposively* manipulates the research context in some way" (Teele, 2014, p. 118; italics my own).

277). Metcalf and Crawford have pointed out the missing focus on human subjects in big data science (2016). While databases with subject data can be re-combined to formulate pictures that are much more invasive than the initial study might be, these interventions are not considered human subject research due to current formulations of what constitutes human subject research (Metcalf & Crawford, 2016). This point has also been made in a recent report by the Ada Lovelace Institute, in which the authors write that:

> The current role, scope, and function of most academic and corporate RECs [*red.* research ethics committees] are insufficient for the myriad of ethical challenges that AI and data science research can pose. For example, the scope of REC reviews is traditionally only on research involving human subjects. This means that the many AI and data science projects that are not considered a form of direct intervention in the body or life of an individual human subject are exempt from many research ethics review processes (2022, p. 8).[25]

In light of these limitations, Resseguier and Ufert have argued in favor of adapting current research ethics standards and mechanisms to better asses scientific AI field research (2023). These discussions point to AI field research challenging existing research ethics.

The ethics of field experimentation has also been the topic of a recent debate in the political and social science. In recent decades, these disciplines have increasingly moved toward field experimentation (experimental interventions outside controlled laboratory environments) as a prominent research methodology. However, this move beyond the lab has prompted questions of ethics and growing calls by scholars that these social and political field experiments are insufficiently ethically regulated (Teele, 2014; Humphreys, 2015; Desposato, 2015; MacKay, 2018; Whitfield, 2019; Beerbohm et al., 2020; McDermott & Hatemi, 2020; Phillips, 2021). For example, McDermott and Hatemi note about political field experiments that:

> We have somehow entered into a Wild West where anything goes when it takes place in the public sphere in large populations, while small controlled laboratory experiments must follow established guidelines" (McDermott & Hatemi, 2020, p. 30019)

Another gap concerns that these political field experiments might bring about unique or different ethical concerns that existing research ethics frameworks do not capture. For example, Beerbohm and colleagues have argued that political field

[25] With exempt here, the report means in an institutional or legal sense.

experiments may undermine 'political equality' between citizens (2020). McDermott and Hatemi have argued that social and political field experiments may harm entire societies, which current research ethics frameworks cannot account for since they focus on harm against the individual (2020). Instead, they have suggested 'respect for societies' as an action-guiding and protecting principle in the design and execution of social and political experiments within society.[26] Similarly, Whitfield has argued that political field experimentation involving human subjects shares the capacity to commit 'interpersonal' and 'diffuse' wrongs with biomedical research but that only political field experiments may bring about wrongs against 'collectives.' Such 'collective wrongs' undermine the decision-making capacity and rights of *groups* rather than individuals, particularly by undermining underlying values such as 'sovereignty' ("the right of states to noninterference in their internal affairs"), 'subsidiarity' ("the rights and authorities of states devolve to substate units and organizations") and 'association' ("the rights of individuals to form associations") (Whitfield, 2019, p.533). Specific examples of political field experiments that might undermine, for example, sovereignty are those that "involve intentionally violating foreign laws" (Whitfield, 2019, p.533), as he claims was the case in (Fried et al., 2010).[27] [28]

---

[26] It remains unclear, however, in McDermott and Hatemi's argument what this principle entails, what would satisfy it, and to what extent it would cover a theoretical deficiency, for example, why 'respect for persons' as a principle is distinct from the societal cumulation of 'respect of persons'. However, McDermott and Hatemi invoke examples that exemplify individual moral concerns, such as not consenting to research participation, or due to the experiment changing features of the world that individuals generally have rights against, such as the non-target population being exposed to increased risks to their welfare. Additionally, it remains unclear how their principle of respect for society relates to those conditions where an overall society may be considered unjust, and the purpose of the research intervention is to (help) solve said injustices.

[27] Whitfield mentions he does not wish to suggest that "lab experiments, semi-, or nonexperimental methods cannot in principle risk similar impacts" as political field experiments (2019, p.536). This somewhat undermines his appeal for a separate research ethics of political field experiments. Instead, it seems rather to suggest that research ethics, in general, should be more cognizant of the specific impacts Withfield draws attention to, which may be caused by political field experiments but also by other kinds of research.

[28] Mackay (2023) has criticized these accounts, noting that, while valuable, they are building on a limited and outdated image of research ethics, for example, by only mentioning the Belmont Report (1978). He notes that: "discussions of the ethics of clinical research have moved beyond this, refining interpretations of the principles and applications found in Belmont, contesting these interpretations, and developing new concepts and principles to evaluate clinical research protocols" (Mackay, 2023, p.3).

While the current debates within scientific research have rightfully drawn attention to both the ethics of field research and that current research ethics guidelines do not capture the full range of moral issues various research fields encounter, they leave a significant gap: they have been primarily focused on the expansion or adaption of research ethics to *scientific* research outside the laboratory. However, much non-scientific real-world research also lacks research ethics governance. Nevertheless, this problem has been only limitedly discussed. I will expand on this in the next section.

3.3. Scientific Research Ethics Exceptionalism

Scholars have increasingly pointed out that research as a domain of human activity seems much more stringently regulated than many seemingly identical activities. The fact that scientists who want to conduct interviews for research purposes need to get approval from an ethical review board while journalists do not is one example of this. However, we find such examples anywhere where research and non-research activities overlap, whether fishing, urban planning, traffic, policy-making, sports, or business (Hansson, 2011). It seems that a vast range of ethically salient activities can be conducted as both (part of) research or outside a research setting. In turn, whether or not something is understood to be research determines the 'ethical demands' placed on this activity (Hansson, 2011). Research activities are thus seemingly treated as 'exceptional' (Wilson & Hunter, 2010). They are subject to higher ethical demands than similar human activities not labeled as research (Wilson & Hunter, 2010; Hansson, 2011). This difference is particularly noticeable when it comes to human subject research. Scholars have disagreed over whether this difference is justified (Hansson, 2011; Wilson & Hunter, 2010). This problem is what Hansson has called the 'boundary problem of research ethics': what exactly – if anything – justifies this differential treatment of ethical demands (Hansson, 2011)?

However, one shortcoming of these accounts is that not *all* research is treated exceptionally. Different kinds of research are not held to the same research ethics standards (as a matter of governance) (Miller, 2010; Moffat, 2010). This boundary of ethical demands that Hansson mentions does not run 'around' research; it cuts right through it. While *scientific* or *governmental-funded* research is often held to clear (but perhaps imperfect) research ethical standards and protocols, the same cannot be said for research conducted by many public and private parties. For example, in the context of online corporate A/B testing, Polonioli and colleagues argue that:

> "The use of human subjects in research that is not federally or publicly funded—such as in the case of privately funded A/B testing, often affecting millions of potentially unaware people—has remained unregulated" (Polonioli et al., 2023, p. 669).

Similarly, commenting on corporate experimentation on online worker platforms in the Financial Times, Rahman writes:

> "The problem is not experimentation in itself, which can be useful to help companies make data-driven decisions. It is that most do not have any internal or external mechanisms to ensure that experiments are clearly beneficial to their users, as well as themselves. Countries also lack strong regulatory frameworks to govern how organizations use online experiments and the spillover effects they can have. Without guardrails, the consequences of unregulated experimentation can be disastrous for everyone" (Rahman, 2024).

This absence is particularly pronounced in contrast to scientific research and other publicly-funded research in which research ethics governance is firmly established. Alternatively, Calo writes that:

> "Any academic researcher who would conduct experiments involving people is obligated to comply with robust ethical principles and guidelines for the protection of human subjects, even if the purpose of the experiment is to benefit those people or society… But a private company that would conduct experiments involving thousands of consumers using the same basic techniques, facilities, and personnel faces no such obligations, even where the purpose is to profit at the expense of the research subject" (Calo, 2013, p.101).

Privately conducted research, thus, is not subject to the same research ethics governance as scientific or publicly-funded research (Moffat, 2010).

To clarify, this is not merely a phenomenon of *corporate* research but also includes research conducted by governmental or public parties such as law enforcement. In the context of testing new technologies for migration management in border zones, Molnar writes that:

> "All this experimentation occurs in a space that is largely unregulated, with weak oversight and governance mechanisms, driven by the private sector innovation" (2020, p. 34).

Alternatively, in 2020, Amnesty International called on the Dutch government to end dangerous police experiments with mass surveillance, citing human rights abuses (2020). In their report, Amnesty outlines the dangers of 'predictive policing,' a method that uses mathematical models to assess the likelihood of a criminal offense

being committed by a specific individual or at a particular location (Amnesty, 2020; Susser, 2021). Amnesty writes:

> "A comprehensive policy and legal framework for regulation and oversight is yet to be introduced. Meanwhile, the police are running several experimental predictive policing projects under the premise that the existing legal framework sufficiently regulates their use of algorithmic models and big data predictions" (2020, p.14).

These accounts indicate that various corporate and public research forms are often unregulated or comparatively less regulated than their scientific or publicly funded counterparts.

Naturally, companies or public institutions can integrate ethics into their operations through ethics committees, codes of ethics, or internal guidelines. However, at the core, these practices often amount to self-regulation. Its value notwithstanding, what separates these practices from research ethics governance in scientific research on at least three counts is that (1) it is externally imposed, (2) it can be mandatory, and (3) it can hold actual sway over whether the research is conducted or can continue. Going forward, I will refer to these two categories as ethically regulated research and ethically unregulated research when describing those kinds of research bound in some sense by externally imposed and mandatory ethical regulations and those that are not.

There have been some calls in the literature to extend ethical guidelines to various areas of ethically unregulated research. Recently, various scholars have discussed the ethics of corporate social media A/B testing and have called to extend current ethical regulations – such as institutional review boards – to these practices (Kramer et al., 2014; Grimmelman, 2015; Jouhki et al., 2016; Benbunan-Fich, 2017; Wood, 2020; Polonioli et al., 2023). Svensson and Hansson have discussed extending research ethics guidelines to traffic experiments (2007). Zimmerman has used research ethics guidelines from psychology to analyze experimental smart city interventions (2023). Taylor has argued for research ethics in urban experimentation, arguing that under technological urban experimentation, research subject populations suffer a problematic 'power asymmetry' and are "disempowered with respect to knowledge, understanding, and agency" (2020, p. 1909).

The point of this chapter differs from those of these scholars in at least two meaningful ways. First, the focus of this paper does not focus on a specific form of testing under real-world conditions, such as A/B tests. Instead, it takes a broader, comprehensive perspective to include various forms of research and development with technology under real-world conditions to make a broader case for harmoniza-

tion, similar to how I have done in Chapter 2. Second, this scholarship has not significantly engaged with the idea that this differential treatment can be justified if meaningful differences exist. While authors such as Hansson (2011) and Wilson and Hunter (2010) have questioned whether scientific research needs more stringent ethical regulation compared to non-research activities of similar risk, there has been limited scholarly attention to examining the justification of different ethical demands between scientific and non-scientific research practices, both in a moral sense and in a governance sense. Hence, in the next section, I examine this idea in the sense of research ethics governance. I will argue that there are no meaningful conceptual differences between these two categories of research that justify the current different research ethics regulatory demands.

3.4. Arguments against meaningful conceptual differences

In this section, I will argue that different ethical regulation for different forms of research is problematic since no meaningful conceptual differences exist between currently ethically unregulated and ethically regulated research. I will examine four possible reasons that might justify why we place research ethics regulatory demands on certain types of research but not on others. I examine the following four (non-exhaustive) reasons: (1) the ability or potential to be harmful, (2) their environmental research conditions, (3) the goal of the research, and (4) the role of the researcher. I will subsequently reject these as sound justificatory reasons.

First, an argument to justify the status quo could be to claim that all those forms of currently unregulated research are less risky than those currently regulated research (Wilson & Hunter, 2010).[29] Introducing research ethics governance would, in that case, amount to disproportionate over-regulation of research that is not (likely to) cause any harm. However, this position is unconvincing. While it may be true that *some* unregulated research with human subjects is not risky, there are plenty of examples of research that are risky or have caused harm, such as an experimental self-driving vehicle operated by Uber killing a pedestrian (Stilgoe, 2020), the Dutch police violated human rights with their predictive policing experiments (Amnesty, 2020), and Uber 'reshaping' the sense of autonomy of platform workers (Rahman et

[29] In this paper, I hold risk to mean the likelihood of harm (Maheshwari & Nyholm, 2022). Additionally, I define harm not merely in a physical sense but as any 'wrongful setback' of an 'interest', such as the violation of a right (Feinberg, 1984).

al., 2023). In other words, independent of their current ethical regulatory status, ethically unregulated real-world experiments with digital technology can carry some risk. This is the same for currently ethically regulated research; some are more risky than others. We find examples of risky and non-risky research in ethically regulated and unregulated research.[30] Thus, there should be no difference between the regulatory status of these two categories of research.

To clarify, I do not claim that ethically unregulated research is *more* risky than ethically regulated research based on its regulatory status or that regulation will necessarily prevent these abuses. As Wilson and Hunter point out, while cases of research risk, abuse, or harm "do provide prima facie evidence that unregulated research *can* be abused […] they do not show that regulation *will* prevent these abuses" (Wilson & Hunter, 2010, p.49; italics my own). Merely having ethical regulation does not mean that researchers will keep to it. In order to make this strong argument for ethical regulation of currently unregulated research, I would either have to demonstrate that (1) there is a direct relation between an absence of research ethics governance and concrete research abuses or that (2) research ethics governance do minimize risks and realize participant safety. However, these are empirical claims, and not only are they outside the scope of this paper, but little empirical evidence exists to support them (Bean, 2010). Instead, the argument above has been limited to claiming that no apparent difference exists in the risk for harm between ethically unregulated and regulated forms of research.

Second, in some cases, the environment in which research is conducted can be argued to determine whether or not we hold research to particular moral norms. For example, Spicker has argued that research conducted in the public sphere does not require consent since participants do not have a right to control their data in the public sphere to begin with (2011). However, considering real-world research, there are no seeming relevant differences between the environmental conditions under which this research is conducted since both regulated and unregulated research can be conducted under real-world conditions. There might be strong arguments that moral obligations can be waived in given research contexts, like obtaining informed consent in the public sphere. However, this is an argument about prioritizing or waiving two competing rights - for example, the right to conduct research and the

[30] That is not to say that over-regulation and under-regulation are not serious issues that can both have ethical implications. Rather, this would not be an argument against harmonizing research ethics protections.

right not to be researched (Traianou & Hammersley, 2021) - based on salient features of the research environment. However, it does not justify placing little research ethics regulatory demands on real-world research to begin with. For example, while Spicker argues that obtaining informed consent is not necessary for studying public actions, he still holds that researchers are still obligated to respect the general rights of those involved in their research and to ensure appropriate safeguards are in place (Spicker, 2011). What these exact demands are could differ. It seems reasonable to some degree to adjust the ethical demands of research ethics guidelines to the challenges of the research field it aims to guide.

Thirdly, the nature of the research goals between these two categories also does not justify unequal research ethics regulatory demands. There can be apparent differences between the goals of various forms of research, such as solving social challenges, developing commercial products, or 'mere' academic curiosity. However, it seems unconvincing that these goals, per default, justify placing different research ethics regulatory demands since both currently ethically unregulated and regulated research can be conducted for economic purposes, solving social challenges, or satisfying academic curiosity. It seems more reasonable to argue - in line with current practice – that different public or corporate research goals, aims, or needs can present particular reasons that override existing demands, especially when research protections are generally taken not to be absolute and can be overridden by outweighing considerations.

Fourth, do the different professional roles give meaningful conceptual reasons to treat these two categories as distinct? Different professions are often assigned specific ethical obligations associated with their professional role (Wilson & Hunter, 2010). Currently, in a governance sense, we currently hold scientific and publicly funded researchers in their behavior toward their research participants to particular ethical demands. In contrast, we do not hold the same research ethics governance demands to researchers who fall outside of this scope. However, the distinction between a researcher in a corporate lab and a university lab is not as straightforward as the different professional roles between researchers, teachers, lawyers, doctors, nurses, engineers, etc. To what extent do the researchers who are or are not regulated have substantively different professional roles? What reasons exist for treating these professional roles as researchers substantially differently regarding ethical demands?

One reason that might justify holding scientists to higher standards would be safeguarding the public's trust in scientific research as a social good (Wilson & Hunter, 2010). However, this is an empirical claim. It is not clear that the public

trust in science is affected by the current higher ethical demands in the first place (Wilson & Hunter, 2010). However, assuming there is empirical data to back up a casual connection between research ethics regulation and public trust, does the public's trust in the research conducted by non-scientific researchers need not be safeguarded either? Corporations, for example, also seem to benefit from public trust in their research to succeed in the market.

Alternatively, scientific or publicly-funded researchers might derive their professional obligations from conducting research with public financial support. Consequently, they might consider they have a particular responsibility to the public, for example, by helping to improve society and not harm it in the process. However, as I pointed out earlier, when public organizations conduct research, such as policy experiments, they are not necessarily held to research ethics regulations despite drawing from public funds. This might suggest we are mistaken in not holding them to the same standards. Additionally, while corporate researchers might not depend on public funds (although sometimes on public investments), they still benefit from society in other ways when conducting real-world research, for example, from people interacting with their system or by using public facilities or infrastructures that public funds maintain. Thus, the nature of public funding alone seems insufficient to justify an absence of research ethics governance.

Additionally, I am unsatisfied with the idea that we should *only* extend particular ethical demands to those professional roles that – through historical coincidence – came to view themselves as a particular profession or vocation with specific responsibilities. I am unsatisfied with this because it would mean we distribute research ethics regulatory demands based on the willingness of a particular professional community, placing the highest research ethics regulatory demands on the most willing. Of course, this does not mean all researchers might be bound by the same moral standards. However, as a matter of research ethics governance, this opens the door to rejecting externally imposed regulation based on the argument that it falls outside their professional role's scope. However, professional roles are not static; they evolve over time. The first developments of modern research ethics in the shape of guidelines outlining the responsibilities of medical scientists were also not readily adopted within the professional community. Equally, researchers in public institutions or corporations might feel that their professional duties do not include the demands of research ethics governance. Nevertheless, we might still have reason to subject them to particular regulatory demands. Moreover, their professional role and associated obligations might change accordingly over time.

Arguably, this point remains somewhat inconclusive. I am sympathetic to the idea that different professional research contexts might present different overriding reasons for the professional obligations of researchers and that these obligations should not be taken to be absolute and can be overriding in the face of sufficient reason. However, it is unclear to me how an appeal to the professional role and duties of a, for example, corporate researcher vis-à-vis a public researcher could be used to argue that research ethics should not be harmonized in the first place. Further work is necessary, however, to work this out further and, for example, ground the professional obligations of researchers in their activities as researchers first and on their employers or funding source second (as is now the case in a governance sense).

This, however, is beyond the scope of this paper. Instead, in the next section, I sidestep the question about whether the professional roles of different researchers bring about sufficient reasons to treat them to different ethics regulatory standards, and argue that the current situation of placing different research ethics regulatory demands on real-world research is problematic because of potential regulatory avoidance of ethical demands, which comes at the expense of those whom these demands aim to help protect. I will elaborate on this in the next section.

3.5. The argument against regulatory arbitrage

In this section, I argue why placing different research ethics regulatory demands on various research practices can be problematic. That is because it creates the opportunity for what I will call research ethics regulatory arbitrage: avoiding particular regulatory demands by placing research activities outside the regulatory scope of the research ethics governance a research project is subject to.[31] This is a problem since this avoidance comes at the expense of those these regulatory demands aim to help protect.

Modern research often involves collaborations between academic, public, and private entities. As Colona argues, modern scientific research "is conducted by a wide variety of actors, including private entities, ranging from small startups to powerful tech giants, as well as governmental and non-profit organizations" (2023, p.1-2). The same is increasingly the case for AI research (Ada Lovelace Institute, 2022). Real-world experiments with emerging technology are often conducted in

[31] This section will not discuss why and how this differential treatment arose.

transdisciplinary teams and organized in so-called 'real-world laboratories' or 'living labs,' where academic, public, and private parties collaborate to develop and test digital technologies or policy interventions in real-world settings (Ansell & Bartenberger, 2016; idem, 2017).

The current differential research ethical demands are problematic given the collaborative nature of modern research practice because they create a problem of different regulatory demands or ethical standards within a collaborative research project (Ada Lovelace Institute, 2022). As such, research parties can avoid taking responsibility for certain research practices by placing them outside the scope of their responsibilities. Since regulatory demands often end at institutional funding boundaries, this opens up possibilities for research ethics 'regulatory arbitrage' (Colonna, 2023). I follow Colonna in defining regulatory arbitrage as "an avoidance strategy of regulation that is exercised because of a regulatory *inconsistency*" (2023, p.2, italics my own). Research ethics arbitrage would be a subset of regulatory arbitrage focused on the avoidance of research ethics regulations. When one category of research is regulated by different ethical demands compared to another, it might be in the interest of those who do not want to adhere to particular burdens of ethical regulation to place or conduct (part of) their research out of the scope of their research ethics governance. Data for which scientific researchers might not get research ethics approval could be collected by industry partners operating outside the scope of institutional review boards. At that point, scientific researchers would be working with 'existing' data, which would not prompt the need for ethical review. This would allow academic researchers or transdisciplinary collaborations to circumvent research ethics governance (Grimmelman, 2015).

An (in)famous example is the Facebook Emotional Contagion study, in which researchers at Facebook, in collaboration with Cornell University, researched emotional contagion through social networks (Kramer et al., 2014). The study was conducted in the following way. Researchers at Facebook would alter the amount of positive or negative posts on the news feeds of certain users to see whether their subsequent posts were affected by this exposure. The 'emotional contagion' then refers to the question of whether exposure to either positive or negative posts would change the emotional state of the user's post positively or negatively. This manipulation resulted in a large set of data that the Cornell researchers were granted access to by Facebook (Flick, 2016). The study generated substantial controversy. Facebook users were unaware that their feed was manipulated, did not give consent to their participation, and the company did not seek ethical review. However, their studies

were aimed at manipulating people's emotions. The Cornell University researchers had sought IRB approval, but since data collection was technically done independently from Cornell before their involvement – and they were essentially working with existing data – the IRB judged that no review was necessary (Flick, 2016). I do not wish to suggest that the Cornell University researchers intended to circumvent IRB approval. However, the exact mechanism can be used intentionally to do just so.

In other words, an uneven field of ethical demands across different research domains leaves space for avoiding these demands. This comes at the expense of those whom these demands aim to help protect. This provides a strong case for the current situation to be amended. Research ethics governance can thus benefit current unregulated real-world experimentation with digital technologies by promoting collaboration and harmonization across disciplines and industries. However, there are challenges to this goal. I will discuss these briefly in the next section.

3.6. Going forward: towards a research ethics of real-world research

In the previous two sections, I have argued that the current situation of different research ethics regulatory demands between different categories of research is problematic since (1) there are no clear, meaningful justifying reasons to place different demands on currently regulated and unregulated research and (2) the current inconsistent situation of unequal demands can be exploited by circumventing ethical demands at the costs of person's protection. However, both arguments can swing in two ways. They can be used to justify reducing the regulatory burden on currently regulated research to bring it closer to unregulated research practice, or they can be used to increase the regulatory burden on currently unregulated research. As Grimmelman argues:

> "To the extent that the Common Rule reflects a consensus about academic research on social media users, it should extend also to corporate research on social media users, because the ethical argument for regulating the latter is at least as strong as the argument for regulating the former … But if corporate social media experiments do not need to worry about informed consent or ethical oversight, we should be having a conversation about exempting academics, too." (2015, p. 254).

While I agree with this, I think a stronger case can be made to bring currently unregulated research more in line with currently regulated research. However, this is outside the scope of this paper (see, for example, Hansson (2011) for an argument in

favor of this position). Instead, assuming that harmonizing ethical demands means higher ethical demands for ethically unregulated research domains, this section will briefly discuss several opportunities and challenges to adapting existing research ethics governance to real-world research.

Polonioli and colleagues have discussed the benefits and challenges of several mechanisms for a 'soft ethics' framework for controlled A/B testing, such as (internal) institutional review boards (IRBs) and questions that aid reflection on the ethics of A/B testing (Polonioli et al., 2023). They rightfully point out that:

> "Companies should not be left alone in trying to elevate their standards of ethical experimentation. Engineers and developers often involved in experiments are not systematically trained in ethics, may perceive ethical considerations as unnecessary red tape, and need to grapple with unavoidable conflicts of interests due to the close link between business and science. To foster compliance with ethical principles, companies need to be properly educated, governed, and incentivized" (Polonioli et al., 2023, p. 679).

However, as they admit, a shortcoming of their approach is that they largely hinge on companies' and institutions' *voluntary* adoption of these recommendations.

Whether self-regulation is sufficient has been called into question. As Grimmelman argues:

> "The history of privacy protections shows that self-regulation without corresponding regulatory oversight is a cruel joke at the expense of users. Unless we start from a place in which social media research is subject to ethical limits, we will not get there. If companies like Facebook and OkCupid know that they must deal with the Common Rule, they will have every incentive to work constructively to fix its imperfections. If they believe that they fall outside it, they will fight tooth and nail to continue in their unregulated ethical free-fire zone" (2015, 259).

Some empirical evidence supports this. A study by Stahl et al. which focused on the adoption of Responsible Research and Innovation (RRI) practices in the European ICT industry, concluded that:

> "Innovators recognise some of the ethical and societal concerns associated with their [research and development] activities but their approach is often piecemeal; primary focus is upon the most immediate issues and on legal compliance, to the detriment of broader societal issues and wider challenges" (Stahl et al., 2019).

Alternatively, while most researchers can grasp the implications of their research since their behavior is being evaluated, there is an incentive to evaluate it over-

favorably (Wilkinson, 2010). This paints a compelling argument for independent research ethics governance, such as review boards, external audits, or legal compliance.

Here, inspiration can be drawn from the EU AI Act and regulatory sandboxes (Ranchordas, 2021; Buocz et al., 2023). The AI Act is a piece of European legislation seeking to regulate the development and deployment of artificial intelligence systems within the European Union. Based on a pre-established risk profile of an AI system linked to its intended purpose, the AI Act makes market approval conditional to specific requirements. However, wishing not to stifle research and innovation, the European AI Act offers exemptions to its regulation for AI research (Colonna, 2023). Given that certain conditions are met, real-world testing with high-risk AI within and outside EU-sanctioned regulatory sandboxes is allowed. Madiega and Van De Pol define regulatory sandboxes as:

> "Regulatory tools allowing businesses to test and experiment with new and innovative products, services or business under the supervision of a regulator for a limited period of time" (2022, p.2).

According to AI Act, this 'supervision' should be done by:

> "A competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real-world conditions, an innovative AI system, pursuant to a sandbox plan for a limited time under regulatory supervision" (Article 3, 55).

This same 'competent authority':

> "shall provide, as appropriate, guidance, supervision and support within the AI regulatory sandbox with a view to identifying risks, in particular to fundamental rights, health and safety, testing, mitigation measures, and their effectiveness in relation to the obligations and requirements of this Regulation and, where relevant, other Union and national law supervised within the sandbox" (Article 57, 6).

Real-world experimentation outside the regulatory sandbox is also possible, given that various conditions are met (Article 60), such as submitting "a real-world testing plan" (4a), limitations on the period for which the testing can be done (4f), protections for subjects "belonging to vulnerable groups" (4g), informed consent for subjects (4i), and ensuring that "the predictions, recommendations or decisions of the AI system can be effectively reversed and disregarded" (4k). Placing similar demands

on ethically unregulated research could be a step towards harmonizing the regulatory demands placed on research.

The benefit of these conditions is that meeting them is necessary for market entry. Companies, governments, and research institutions thus have a financial and legal incentive to abide by them if they want to develop and, eventually, deploy, in this example, a particular AI system. Governments or oversight authorities could use a similar structure to regulate under what conditions researchers could access public space for real-world experimentation. However, they are also quite substantial requirements. They place significant demands on oversight authorities to enact them and researchers to abide by them, thus raising questions of proportionality to regulate various forms of real-world experimentation. Additionally, these conditions would need further conceptual specification (what exactly constitutes 'additional safeguards for vulnerable groups'?), sensitivity to undermining circumstances (what is the value of consent when a participant has limited alternative options available?), and overriding reasons (under what conditions should these demands be allowed to be overridden?).

3.7. Conclusion

In this paper, I have drawn attention to inconsistencies in the current ethical regulation between various categories of real-world research. Testing technologies under real-world conditions is widespread, yet the ethical issues it raises are often neglected. I have argued that these inconsistencies in ethical demands and protections are problematic. No apparent meaningful difference warrants this inconsistency in research ethics governance, and it creates the possibility of regulatory evasion at the cost of public protection. I have grounded my argument in several larger scholarly debates on the limits of current research ethics governance in the face of novel technologies and research formats. I contribute to these debates by drawing attention to a new area needing research ethics reforms. I have briefly discussed several ways forward by drawing inspiration from the AI Act's current regulation on real-world testing of high-risk AI, yet pointed out that these approaches bring about new problems of their own.

Prologue to Chapter 4

In the previous chapter, I analyzed the lack of research ethical governance for real-world research. I argued that this lack is unjustified by morally significant differences between those practices, at least when considering four initially plausible but fallacious differences, and that we should thus develop research ethics governance for real-world research. While I argued to reduce inconsistent research ethics regulatory demands between research domains, I did not argue that this governance necessarily needs to be based on the same norms and principles dominant in the current research ethics paradigm. I turn to this concern in the next chapter.

In Chapter 4[32], I analyze whether real-world research can uphold paradigmatic research ethics norms such as informed consent, the right to withdraw, just distribution of risk, benefits, and research participants. I do not presuppose that the moral obligations that researchers have to participants or bystanders must necessarily be the same as to those in scientific human subject research. Rather, I note that these paradigmatic norms have been used to evaluate relevant examples of real-world research and ask – if we would assume that we should hold real-world research to these paradigmatic norms – whether such norms can be upheld in real-world research. Specifically, I ask:

> RQ4: *What challenges emerge in applying paradigmatic research ethics norms to real-world research?*

I identify and focus on one challenge: a problem of identification. Because real-world research is conducted in natural or uncontrolled environments, it involves significant uncertainty and difficulty in identifying which persons they subject to the research. This practical concern undermines an assumption on which many paradigmatic research ethics norms rest: that we can identify those individuals to whom researchers might owe particular obligations. This would not mean that these obligations are not owed as a matter of morality, but simply that they would be difficult or impossible to uphold in current practice.

---

[32] Submitted as: "Experiments without Borders: Research Ethics and the Identification Problem of Real-World Research" by Joost Mollen.

4. The Identification Problem of Real-World Research

Abstract

Research with technologies under real-world conditions has become an increasingly pervasive phenomenon, yet often lacks ethical governance. In response, scholars have applied paradigmatic research ethics norms to analyze and evaluate this real-world research. Such ethical norms include but are not limited to obtaining informed consent, ensuring the right to withdraw, weighing the research's benefits and risks to participants, compensating or remedying research-related harm, containing adverse consequences, debriefing participants, ensuring a just distribution in participant selection, and offering extra protections to vulnerable participants. However, these norms assume that research participants are known or can be identified, which is often not the case in real-world research. I will refer to this problem as the identification problem. If we maintain that these paradigmatic norms apply to real-world research, then the identification problem presents a severe moral challenge, and overriding reasons must be identified. In this paper, I connect a fragmented debate on this issue and offer a descriptive account of the origin, scope, and consequences of this phenomenon. I discuss the implications and significance of this problem to the growing field of real-world research and outline ways forward. Without otherwise justificatory reasons, the identification problem renders real-world experiments morally impermissible. This prompts further scholarly reflection on developing research ethics guidelines and governance structures for real-world research practices.

4.1. Introduction

As I have outlined in Chapters 2 and 3, there has been an increase in attention to research formats that are conducted under real-world conditions, intervene in complex but 'natural' socio-technical environments, and, depending on the intervention, operate close to a public engaged in daily life. To reiterate, examples of these phenomena include political and social field experiments (Humphreys, 2015; Desposato, 2018; Whitfield, 2019; McDermott & Hatemi, 2020; Phillips, 2021), online digital experimentation such as A/B tests on social media (Grimmelman, 2015; Benbunan-Fich, 2017; Poloni et al., 2023), field studies on human-robot interaction in public spaces (Mollen et al., 2023a), and experiments organized in pilot projects and living labs for social innovation and technology development (Ansell & Bartenberger, 2016; Taylor, 2021; Pfotenhauer et al., 2022). Examples of the latter include tests in public with, for example, self-driving cars (Fehlmann, 2019; Marres, 2020; Stilgoe, 2020), predictive police technology (Amnesty, 2020; Susser, 2021), and crowd control technologies such as mood-altering streetlights (Galic, 2019) and monitoring drones (Barmpounakis & Geroliminis, 2020).

There has also been increasing scholarly attention paid to the normative dimension of real-world research. Scholars have commented on the absence of proper regulatory control and ethical regulation (Taylor, 2021), on investigators not complying with established research ethics guidelines (McDermott & Hatemi, 2020), on the friction between established research ethics norms, such as the right to withdraw, and particular practices in real-world research (Mollen, 2023b), the need to include the voices of those affected by real-world research in ethical considerations (Fiesler & Proferes, 2018), or the inability of existing ethical guidelines to accurately capture the full range of moral concerns regarding real-world experimentation within a specific scientific domain such as big-data research or political field experimentation (Metcalf & Crawford, 2016; Whitfield, 2019).[33]

However, one challenge that has received less explicit attention is that research formats conducted under real-world conditions face significant uncertainty about whom the research involves and affects. For example, in the context of smart city experiments, Zimmermann writes that the 'interconnectedness' involved "makes it

[33] Specific ethical challenges that real-world experiments have brought about include the absence of public awareness about experiments taking place and the manipulation of the public (McDermott & Hatemi, 2020), causing physical harm (Stilgoe, 2020), violating human rights (Amnesty, 2020) and the imposition of 'dominating' risk (Maheshwari & Nyholm, 2022).

difficult to determine the boundaries of the socio-technical system, to isolate effects and to limit interventions ... System boundaries become increasingly blurred, leading to an unclear reach of interventions" (Zimmermann, 2023, p. 54). Kroes notes that when we conduct experiments on socio-technical systems, "there is the problem of where to draw the line between the experimental system and its environment" (2016, p. 642). Additionally, de Graeff and colleagues note that community engagement is understood to be an important element of the design and launch of field trials with gene drive technologies (de Graeff et al., 2023). However, as de Graeff and colleagues point out, "disagreement and lack of clarity exist about how this "community" should be defined and delineated" (de Graeff et al., 2023, p. 601). This creates problems in identifying community members to whom a researcher might hold particular obligations if they are to be included in community engagement efforts (de Graeff et al., 2023).

In this paper, I capture this phenomenon as the *identification problem*. I argue that the identification problem plagues various interventions across research domains and challenges the application of certain paradigmatic research ethics standards for ethical and responsible research conduct which presuppose an investigator knowing who is involved or affected by a research intervention. When investigators do not have this knowledge, those norms are impractical or impossible to uphold. Without overriding reasons, these research formats then face a moral problem.

While different authors have touched upon the identification problem, several gaps persist. First, the identification problem itself is rarely explored but is taken for granted as a necessary feature of real-world research. Second, the scope of the identification problem is underestimated. Authors discussing the problem have generally focused on informed consent (Schinzinger & Martin, 1983; Van de Poel, 2016). Instead, I argue it affects a far broader range of research ethics norms, such as the right to withdraw, weighing the research's benefits and risks, compensating or remedying research-related harm, containing adverse consequences, debriefing participants, ensuring just distribution of research participants, and protecting vulnerable demographics. Third, the identification problem has far-reaching implications for our efforts to conduct morally permissible research 'in the wild' since, if we take these research ethics norms seriously, it either means that much of real-world research cannot uphold these norms or that we have to redesign our real-world research practices to make sure they can. Alternatively, if we do not take these research ethics norms seriously, this prompts the question of what norms real-world research should uphold in that case. This has been under-explored. This paper aims

to connect these fragmented discussions by capturing the underlying phenomenon and providing a descriptive account of its origin and ethical consequences.

This paper proceeds as follows. In the first section, I will sketch the identification problem. I will argue that researchers are expected to uphold particular research ethics principles and norms towards the public that are difficult to uphold when it is unclear who a research intervention potentially involves or affects. Since the aim of this paper is descriptive, I do not argue in favor of any specific research ethics principle or norm for investigators or participants, nor do I seek to elucidate the conditions under which one principle or norm supersedes another. Instead, I argue that if we take these research ethics principles and norms seriously in the context of real-world research, the identification problem poses a practical challenge to their realization. In the second section, I provide various examples of research ethical obligations that the identification problem problematizes. In the third section, I argue that this identification problem arises as a matter of control over the experimental system, the control conditions in which the experiment takes place, and certain control conditions already presented in that environment. Here, I will argue that not every environment in which research is conducted 'naturally' offers the same 'means' of control for identification**.** Finally, I discuss the identification problem's implications for real-world experimentation and outline the need for further academic research into various resolving strategies.

4.2. The Identification Problem

Research ethics, as a field of applied moral philosophy, is concerned with how research ought to be conducted, how researchers ought to behave, and how we ought to treat others in the name of research. It concerns moral judgment on the permissibility or acceptability of research and, for example, which principles and norms researchers should adhere to for their research to be ethical. In defining principles, I follow Beauchamp and Childress's claim that principles are "normative generalizations that guide actions" but which "leave considerable room for judgment in specific cases and that provide substantive guidance for the development of more detailed rules and policies" (Beauchamp & Childress, 1994, p. 38). Examples of such principles include: 'non-maleficence', 'beneficence', 'justice', and 'respect for personal autonomy' (Beauchamp & Childress, 1994). When I refer to norms, I refer to these 'more detailed rules' that researchers ought to follow (Beauchamp & Childress, 1994, p. 38). For example, while a more general principle might guide us to ensure

that research respects a person's autonomy, a more detailed rule (norm) through which we might meet that principle could be to obtain the informed consent of persons before subjecting them to research. Such norms can consequently be codified in research ethics governance, for example, in ethics guidelines.

Many of such research ethics norms require that an investigator knows who a particular research intervention potentially involves or affects. A prominent example of such a norm is obtaining informed consent. The idea of consent is that it protects a person's right to autonomy or self-determination (Gelinas et al., 2016). Consequently, for the research act to be ethically permissible, those affected need to indicate by their consent that they were in a position to choose to participate freely, were sufficiently informed about the study, its purpose, procedure, and potential associated risk, and had the opportunity to decline or refuse (Nijhawan et al., 2013). In order to be able to provide a person with the opportunity to consent, researchers need to be able to identify those persons who are participating or to whom this norm extends. I will refer to such norms as 'identity-based' norms, because they require the identification of an other in order to adhere to them.

However, in many research formats conducted under real-world conditions, upholding these identity-based norms is not always straightforward due to difficulty determining precisely who needs to be included in upholding this norm. This might be a problem for conceptual or practical reasons. A conceptual identification problem arises where there is unclarity or disagreement about who *conceptually* should be considered a research participant, bystander, or part of a particular community the research targets or seeks to engage with. For example, de Graeff and colleagues write that community engagement is held to be important in the design and launch of field trials with gene drive technologies; however, "disagreement and lack of clarity exist about how this "community" should be defined and delineated" (de Graeff et al., 2023, p. 601). De Graeff and colleagues refer to this problem as the boundary problem: "how boundaries of inclusion and exclusion in community engagement should be drawn" (2023, p.608). Aside from disagreement about what constitutes a community member, another conceptual problem can stem from disagreement about what it means to be affected by research. Long notes that the boundary between being affected and unaffected by research is ill-defined (Long, 1983; Hansson, 2006). Another problem could be that the system on which experiments are conducted is conceptually challenging to delineate. For example, Kroes has noted that it is often unclear what can be considered part of a socio-technical system (2016). He writes that when we conduct experiments on socio-technical

systems, "there is the problem of where to draw the line between the experimental system and its environment" (2016, p. 642). Thus, the *conceptual* identification problem concerns conceptual uncertainty about who belongs to the target group of particular obligations.

There are also practical reasons that research subjects cannot be identified in real-world research. Consider, for example, the following case study from the technology ethics literature. In March 2018, Elaine Herzberg was fatally hit by a self-driving vehicle operated by Uber while crossing a four-lane road near Tempe, Arizona (USA) (Fehlmann, 2019; Stilgoe, 2020). The car was part of an experimental fleet of self-driving cars equipped with sensors and cameras, refining their autonomous driving algorithms while driving on public roads. More than just an unfortunate accident, the collision resulted from a poorly designed experiment. The car's contested detection software failed to identify Herzberg as a point of collision, the human monitor in the car was distracted by their smartphone, Uber's standards for test drivers were lower compared to other self-driving car companies, and the public had not been made aware by Uber or the State of Arizona about the experiment (Stilgoe, 2020).

Such a test conducted under real-world conditions makes executing identity-based norms difficult. Before the experiment, it would have been challenging for Uber to determine who would be involved. Due to operating its vehicles on public roads, Uber could not identify who would be or was involved in their experimental car and its associated risk. With 'being involved' in an experiment, I refer to a situation in which the experimental intervention intentionally or unintentionally influences a person or their environment. With the first, I refer to a situation in which an experiment intentionally imposes particular conditions upon a person or their environment for the purpose or as part of the experiment. Thus, this includes manipulations of the subject and the subject's environment performed for research purposes,[34] but also situations in which an experimental intervention unintentionally influences a person or their environment that is not intended to be part of the experiment but subsequently ends up being involved nonetheless, for example, as a side effect of the intervention or the change it causes in an environment. Kimmelman refers to these persons as 'research bystanders' (2020).

[34] This definition is roughly analogous with common conceptions of 'human subject' within research ethics. See, for example, the Common Rule (US HHS 45.CFR.46).

The identification problem can thus be captured as follows:

(1) Investigators have ethical obligations towards specific individuals or groups involved in their research.
(2) In order to meet particular ethical obligations, researchers uphold particular norms that require that specific individuals or groups be identified.
(3) There is research in which specific individuals or groups cannot be identified.
(4) The norms that require that specific individuals or groups be identified cannot be upheld.

Of course, we can question why we should hold real-world research practitioners, which might be non-scientific entities, such as governments, corporations, engineers, hospitals, NGOs, etc., to the research ethical demands of their scientific counterparts. However, I do not claim that the moral obligations that researchers have to participants or bystanders must necessarily be the same as to those in scientific human subject research. Rather, I note that these existing research ethics norms have been used to evaluate relevant examples of real-world research. Scholars have routinely taken paradigmatic research ethics norms, principles, protocols, and governance structures and use this frame to argue for either extending current research ethics norms to various specific examples of real-world research – such as traffic experiments (Svensson & Hansson, 2007), urban experimentation (Taylor, 2021) and smart city interventions (Zimmermann, 2023) – or demonstrate how various real-world research practices violate paradigmatic research ethics norms when evaluated through this frame, or both. For example, scholars have argued how corporate social media A/B tests routinely fail to obtain informed consent from participants, do not inform them about the research, deceive participants, and so require ethical codes and governance structure such as independent review or research ethics committees in order to ameliorate these issues (Grimmelman, 2015; Jouhki et al., 2016; Benbunan-Fich, 2017; Wood, 2020; Polonioli et al., 2023; Weiss, 2024). In this paper, I build on this scholarship and ask - if we would hold real-world research to these paradigmatic norms – whether particular challenges arise due to the nature of real-world research. I argue that one such challenge is the identification problem.

I am not the first to draw attention to this issue. However, when the identification problem has been discussed in the literature, it is often only concerning the notion of informed consent (Long, 1983; Hansson, 2006; Van de Poel, 2016). Informed

consent can be 'problematic', so argues Van de Poel, because "it may be very hard to identify all individuals that are potentially affected by the introduction of a new technology into society (even if it happens only in a part of society) and to ask them for their informed consent" (2016, p. 672). Schinzinger and Martin refer to this problem as the "consenter identification problem" (1983, p.77). This problem captures that the boundary between being affected and un-affected (and thus the target for obtaining consent) is often not very sharp (Long, 1983; Hansson, 2006). As such, Hansson argues, "the persons whose consent should be sought cannot be identified" (2006, p. 152).

However, I argue that the scope of these accounts is too limited. We need to move from a 'consenter identification problem' to a more general 'identification problem.' This is because, as I will argue, aside from informed consent, it would affect many identity-based norms such as the right to withdraw, weighing the research's benefits and risks, compensating or remedying research-related harm, containing adverse consequences, debriefing participants, and ensuring a form of just distribution of research participants and protecting vulnerable demographics, to name a few. In the next section, I will show how the broad scope of the identification problem affects many identity-based norms.

4.3. The Scope of the Identification Problem

In this section, I will argue that the scope of the identification problem extends considerably beyond affecting mere informed consent. It affects many identity-based norms, such as the right to withdraw, weighing the research's benefits and risks, compensating or remedying research-related harm, containing adverse consequences, debriefing participants, ensuring just distribution of research participants, and protecting vulnerable demographics. This is not meant to be a complete list of the practical and moral challenges of the identification problem for investigators. However, I believe it is indicative of its scope.

First, the identification problem poses problems in awarding participants the right to withdraw. Referring to the earlier example of Uber's Self-Driving Experiment, residents or city visitors had little personal control over opting in or out of the risky conditions Uber imposed on them. Like informed consent, the right to withdraw protects a person's right to self-determination. However, instead of offering participants an informed choice, the right to withdraw prescribes that, at any time, participants can retract their consent to participate in research and stop their

participation without being penalized for making this decision (Edwards, 2005; Schaefer & Wertheimer, 2010; Holm, 2011; Mollen, 2023b). There is disagreement about what the right precisely constitutes and to what extent a person should be allowed to withdraw (Schaefer & Wertheimer, 2010). Does this, for example, merely include a person's immediate participation or the data produced during participation (Schaefer & Wertheimer, 2010)? Aside from the problem of identifying exactly to whom this right should be awarded, the unclear scope of a real-world experiment also seems to question what it means to withdraw in such a context. If an investigator cannot identify precisely who belongs in their subject pool, it is also difficult to exclude a person if they wish – or their data after the fact.[35]

Second, research risks are often weighed against the benefits the research might produce for a specific individual or society. These risks can be weighed based on different principles. For example, Hansson distinguishes between 'collectivist' and 'individual' risk-weighing principles (2004). Under the first principle, Hansson argues, "an option is acceptable to the extent that the sum of all individual risks that it gives rise to is outweighed by the sum of all individual benefits that it gives rise to" (Hansson, 2004, p. 146). Under the latter principle, Hansson argues, that "an option is acceptable to the extent that the risk to which each individual is exposed is outweighed by benefits for that same individual" (Hansson, 2004, p. 146). However, would an investigator aim to use either of these principles to weigh the benefits and risks of the real-world experiment they plan to conduct, they will find that neither balance can be made when the exact scope of who the experiment might expose to risk is unknown. So, even in cases where risks can be expected, the identification problem presents a problem in summing these individual risks or weighing these individual risks against individual benefits.

Third, even if no risks are expected, but harm does arise, the identification problem presents a problem after a real-world experiment has been conducted to identify who exactly has been subject to harm. In these cases, the identification problem presents a problem of compensating or remedying harm. It is generally accepted that persons are entitled to some form of 'compensation' when they are the subject of undue research harm (Kerrison, 2012; Van de Poel, 2016). However, fulfilling this demand is challenging if investigators do not know who they have harmed or might still harm after the experiment is concluded. If persons are also

[35] I discussed a similar problem with withdrawing in the context of live-in laboratories in Mollen (2023).

unaware they are experimented upon, they might not be aware that they are being harmed. Alternatively, if participants have been harmed in some way, they might not know the source of this harm. Additionally, harm might arise after the experiment is concluded, for example, due to latent consequences or the combination of seemingly innocent datasets through data analytics (Metcalf & Crawford, 2016). However, if researchers cannot identify who is harmed during or after the experiment, obligations concerning compensating or remedying harm become difficult or impossible.

Fourth, a related problem concerns the spread and containment of adverse consequences. Researchers should contain the spread of risks and harm as much as reasonably possible (Van de Poel, 2016). According to Gross and Krohn, the benefit of the laboratory is that it allows mistakes made during research to be contained in "a special world" where "the costs of trial and error can quickly be forgotten" (Gross & Krohn, 2004, p. 38). When experimentation is conducted within a complex real-world environment in which the experiment's exact reach is unknown or uncertain, containing harm can become more difficult. This depends mainly on the nature of the experiment in question. An extreme example is the lasting side effects of the French nuclear testing in the Algerian Sahara from the late 1950s to the early 1960s, which contaminates the area to this day (Regnault, 2003; International Atomic Energy Agency, 2005; Hennaoui & Nurzhan, 2023). The nuclear program spread fall-out over many neighboring countries and had environmental and health implications for the local population. When France officially acknowledged responsibility for these consequences in 2010 and offered compensation and medical support for those affected, many affected had already passed away or were difficult to identify and hence compensate. Additionally, nomadic communities collected nuclear-contaminated scrap metal from former testing sites, smelting it into radio-active jewelry and kitchen utensils, extending the experiment's impact far beyond its original location (Aljazeera, 2015).

Fifth, even if no risk was expected and no harm materialized during or after the experiment, the identification problem presents an additional challenge to identify persons for debriefing. Debriefing resolves an absence of information, any false beliefs or negative feelings resulting from the study, or awards some kind of benefit the researcher might owe the subject due to their participation (Verbeke et al., 2023). Some ground this need for debriefing on the idea that research participation is a voluntary and free activity, and persons have the right to determine whether they want to support a specific research cause (Sommers & Miller, 2013). Others under-

stand debriefing as 'moral accountability' (Benham, 2008; Verbeke et al., 2023). This 'moral accountability' is based on the idea that participants were deceived by involving them in a study without their knowledge or consent (Benham, 2008). This deception is considered harmful and by debriefing these persons, these harms can be mended (Verbeke et al., 2023). By not debriefing these subjects, then, is to let the deception and harm persist. Under certain circumstances, it is generally permitted that research is conducted on either unaware or deceived participants as long as these persons are debriefed after participation is concluded (Sommers & Miller, 2013). For example, Sommers and Miller argue that some research is "socially valuable minimal risk research in which reasonable persons would not object to their participation" (2013, p. 112). In these cases, they argue that "contrary to current practice, omitting debriefing is ethically acceptable only when debriefing is impracticable, the deception is innocuous, and no reasonable person would object to involvement in the research" (Sommers & Miller, 2013, p. 98). However, this means that when the deception involved in a real-world experiment is not 'innocuous' or is of a kind to which a reasonable person would object, the identification problem presents a challenge to not debriefing the affected participants.

Sixth, an inability to identify whom you affect has consequences for meeting norms that aim to ensure some form of just distribution. Examples of such norms include fair subject selection, protecting 'vulnerable subjects', or ensuring that research in which 'vulnerable subjects' participate is beneficial to them (Van de Poel, 2016). Take, for example, the idea of awarding specific vulnerable demographics additional protections. There is substantive disagreement within the literature about what vulnerability means, what it means for a person or group to be vulnerable, and who falls within this category (Luna, 2009). However, it is generally agreed within research ethics that there is a possibility that particular individuals might be at higher risk of harm than others when both are subject to the same intervention and that investigators should be mindful of this and employ various tools to mitigate these problems (Waern, 2016; Bracken-Roche et al., 2017; González-Duarte et al. 2019). One reason might be that the person in question cannot fully understand the nature, purpose, or expected and unexpected beneficial or adverse outcomes of participating in a particular experiment (Waern, 2016). However, when the exact make-up of the participant pool is unknown, and the boundaries of the experiment are not controlled and mediated, this allows for persons to be involved *indiscriminately* solely based on their proximity to the intervention.

In this section, I have argued that the scope of the identification problem reaches far beyond concerns about informed consent, but instead affects a great range of research ethics norms that can be characterized as identity-based norms. In the next section, I will aim to clarify the origin of the identification problem.

4.4. The Identification Problem and Control

In this section, I will clarify what causes this problem of identification. I will argue that the identification problem primarily arises as a matter of control over three factors: (1) control over the system on which the research is conducted, (2) the control exerted over various aspects of the research environment, and (3) certain control conditions already presented in that environment since not every environment in which research is conducted 'naturally' offers the same 'means' of control for identification**.**

First, one reason the identification problem could arise is that the system that is under study cannot be fully controlled. For example, Kroes argues that socio-technical systems are often complex systems that involve the behavior of many human actors, which makes them difficult to control (2016). Even if it was possible to identify every individual that could be said to be part of this socio-technical system beforehand, the makeup of those involved could change throughout an experiment due to many difficult-to-control events (Kroes, 2016). A related problem concerns the complexity or interconnectedness of the system under study, such as institutions, cities, industries, or nations. Take smart city experiments, which aim to optimize urban processes using technological intervention. Such a socio-technical system might be highly interconnected with other processes, which, according to Zimmermann:

> "makes it difficult to determine the boundaries of the socio-technical system, to isolate effects and to limit interventions …. to a single individual and a device, within a certain unit of an organization or within the target group only" (2023, p. 54).

In these cases, the exact individuals involved are difficult to identify because they are part of a system over which the investigator does not exercise control.

Second, real-world research involves little researcher control over various aspects of the research environment, which is a cause for this identification problem. As mentioned in Chapters 2 and 3, real-world research is conducted in 'everyday' or 'real-life' contexts,' in 'natural' environments, or as 'real-world,' 'field,' 'social' or 'wild' experiments (Kroes, 2016; Ansell & Bartenberger, 2016; 2017; David & Gross,

2019). What these determinators generally refer to is the idea that, in contrast to laboratory or controlled experiments, these experiments are characterized by relaxed, lower degrees or even an absence of control (Harrisson, 2005; Ansell & Bartenberger, 2016; Hansson, 2015,2016; Kroes, 2016). For example, David and Gross write, "real-world experiments are generally less controllable than experiments in a laboratory" (2019). Ansell and Bartenberger speak about control not being the "*raison d'etre*" of real-world 'generative' experiments. Instead, this research is interested in 'contextual' success and is therefore likely to be conducted under real-world conditions (Ansell & Bartenberger, 2016 & 2017). Ansell writes about 'design experiments', which drop "the pretense of being able to fully control social variables" (2012, p. 172). Van de Poel also separates 'social' experimentation from laboratory experimentation based on the notion of control. According to Van de Poel, 'social experimentation' is different from what he calls, "standard scientific experiments" because, among other reasons, social experimentation is a form of 'uncontrolled' experimentation, in the sense that (1) experiments in society can usually not be controlled by investigators and (2) that they do not control variables to find cause-effect relations (2017, p. 64-65).

However, what control means – or is exerted over – is not always explicit in these examples. In the context of experimental practice, the concept of 'controls' often refers to those control mechanisms involved that aim to guarantee the experiment's internal validity, such as randomization, control groups, isolating variables, or certain statistical methodologies (Ansell & Bartenberger, 2017). Framed through this epistemic explanation of control, we can understand the low control of real-world experimentation to refer to a relatively low degree of application of these techniques and the idea that the success of a particular real-world experiment hinges little on the application of these techniques. However, while this might explain that control has some epistemic function, it says little about what control exactly is. Additionally, the definitions mentioned above also seem to claim that real-world experiments themselves are, in some sort, less controllable or uncontrollable by investigators.

Control, in these examples, seems, therefore, better to additionally conceive of as a kind of 'personal control' (Skinner, 1996). With this, I refer to whether an investigator's actions can intentionally produce a desired outcome or prevent undesired ones in relation to the research environment. 'Low' control would then refer to some low degree of probability that there is a relation between the action of an investigator and the research environment.

This kind of control over the research environment is an essential factor in being able or unable to identify the individuals that research involves. If researchers have much control over the environment under which an experimental system is studied, they can more likely identify and mediate who interacts with the intervention and, subsequently, follow particular identity-based norms such as obtaining informed consent or preventing vulnerable individuals from being subject to research.

Third, the severity of the identification problem can vary, however, based on certain control conditions already presented in that environment. We can make a difference between an investigator adding specific 'means of control' to an environment as part of the experimental design to exercise control and the 'means of control' already present in the environment where the research is conducted. It seems that not every environment in which research is conducted 'naturally' offers the same 'means' of control for identification**.**

Consider, for example, the difference between the following four real-world research examples; conducted online, in a private location, in a fixed public place, or throughout a more extensive public area, respectively. First, a collaborative study between Facebook and Cornell University researched emotional contagion through social networks (Flick, 2016; Juhki et al., 2016). Researchers at Facebook would alter the amount of positive or negative posts on the news feeds of specific users in order to see whether their subsequent posts were affected by this exposure (Kramer, 2012; Flick, 2016). Second, Amazon tested a new robot prototype called Digit in their US warehouses to test whether the robot could work safely with humans (Vallance & McCallum, 2023). Third, the Dutch police tested predictive policing technologies in a famous street and nightlife center in Eindhoven (Galic, 2019). Sound from the street was analyzed for signs of aggression, social media activity in the area was analyzed to deduce visitor's mood, and visual data was analyzed with AI technologies to identify specific arm and leg movements that could be an early sign of an incoming fight and light technologies aimed at altering people's mood to calm down a potentially increasing angry crowd (Mollen, 2018). Fourth, between 2017 and 2018, the Dutch police conducted a large-scale experiment with body cameras to test whether it would increase the safety of police agents (Flight, 2019) One hundred police officers were equipped with two types of cameras: one that records only and one that sends live footage directly to a control room. Officers could choose to wear a camera themselves, and they decide when to activate the camera.

All these real-world research examples were conducted in environments where the researchers added little additional means of control. In all four cases, the people

involved – Facebook users, Amazon warehouse workers, nightlife visitors, and the public interacting with the police – operated in an everyday social context. In all these cases, the identification problem might arise. However, it does not seem that each experiment similarly suffers from the identification problem. Theoretically, investigators have complete control over who is involved in an experiment on a digital experiment in online environments. Facebook has, at least in theory, complete control over which profiles are subjected to particular emotionally charged content. In the second example, since this test was conducted within a private Amazon warehouse, to which theoretically only personnel are granted access, it seems that Amazon would have little trouble identifying exactly who they would be involved in this experiment if they chose to do so. However, identifying everyone involved in the two policing experiments on public streets seems much more challenging. The experimental environment offered them little means for control, and the experimental design did not include any through which researchers could realize desirable outcomes regarding identifying who was involved in the experiment. Whether the identification problem is a problem for an experiment that exercises low control over the environment in which it intervenes depends seemingly much on the exact control conditions already present in the environment. However, it would seem worse when researchers have the means of control, yet do not exercise them (either by choice or by negligence).

The practical problem of identification thus arises from a degree of control over a combination of factors: the experimental system under study, the environment in which the experiment takes place, and certain control conditions are already present in that environment. In this next section, I will discuss the implications of the identification problem in real-world experimentation practice.

4.5. Implications of the Identification Problem

The implications of the identification problem for real-world research are significant. The identification problem implies that many investigators conducting real-world research would not be able to uphold paradigmatic research ethics norms; were we to keep them to those norms. This would render much real-world research morally problematic, which is given extra force in the context of their widespread popularity and alleged epistemic benefit.

Of course, this does not necessarily mean that all real-world experiments are morally impermissible by default. Even if we hold them to these paradigmatic

norms, participants' rights are not taken to be absolute, and competing considerations can outweigh them (Spicker, 2011; Gelinas et al., 2016). However, even if we accept particular overriding conditions, this will leave research cases that do not meet these conditions. Additionally, just as the scope of the identification problem exceeds merely a problem of obtaining informed consent, so does a solution to this problem involve more than merely outweighing a single obligation. Instead, overriding reasons ought to be found for as many as the research obligations affected. This could prove troublesome since it involves a potentially long list of affected obligations; no broad scholarly consensus exists on conditions for overriding participant rights, and overriding reasons for specific norms should also be collectively compatible. Aside from moral, there might also be additional legal implications, depending on the context. For example, as I mentioned in Chapter 3, the European AI Act places specific mandatory provisions on testing high-risk AI systems in real-world conditions, including identity-based norms such as obtaining informed consent from participants (see Article 60, 4i).

In order to address this problem, this disconnect between what research ethics we expect researchers to follow – either in a sense of governance or morality – and the realities of current real-world experimentation practice needs to be bridged in order to ensure that real-world research is being conducted in a way that meets the moral obligations owed to involved (vulnerable) participants and bystanders. Several recommendations follow from this disconnect.

First, as I have already argued in Chapter 3, this disconnect prompts further questioning regarding whether we should hold real-world research to the same standards as other forms of more controlled research or whether we should develop a custom research ethics for real-world research to better align with their practice's realities and challenges. The identification problem should, at the very least, be considered as a core challenge for such a research ethics to address.

Second, taking the earlier mentioned research ethics norms (such as informed consent, the right to withdraw, fair participant selection, etc.) seriously, this disconnect foregrounds the need to adapt current real-world research practices to prevent the identification problem from arising if we hold these norms to be important for real-world research as well. This might be challenging, but research can always be designed so that the identification problem does not arise, for example, by placing additional control conditions in the environment under study or conducting the experiment in a laboratory setting. Of course, this might have consequences to what degree we can still consider the research to be 'real-world' research. Additionally,

particular research might be challenging or even impossible to conduct when researchers impose environmental controls to prevent the identification problem from arising. To some scholars, this is a feature, not a bug. Hatemi and McDermott, for example, have argued that ethical real-world research should be difficult and that technology can go a long way in overcoming this problem (2022). They argue:

> "If one can manipulate millions, then one can consent millions. Even if individual consent is difficult, technology allows for a number of ways to inform the public that a large-scale experiment is about to be released. Local, state, and federal governments do this when making public service announcements, including notifications regarding road closures, risk of fire, and Amber alerts. Radio, Internet, billboards, phone notifications, and television do work" (2020, p. 30019).

Third, the disconnect raises questions regarding the governance of real-world research. If paradigmatic research ethics principles and norms are difficult to uphold when conducting real-world research, what conditions should be placed on researchers and real-world researchers regulating how they can and should be conducted? As I argued, in the previous Chapter 3, this governance is currently lacking. While such demands are often placed as part of scientific research, aside from the law, public and private parties lack meaningful governance mechanisms to ensure they conduct ethical real-world research. Private and public parties should be helped in developing these mechanics (Polonioli et al, 2023). For example, ethical review boards could help educate researchers in the public and private sectors who might be unaware of these moral challenges and how to navigate them.

4.6. Conclusion

In this paper, I have provided a descriptive account of what I have called the identification problem. The identification problem arises because upholding particular research ethics norms requires that individual research participants are identified, but these specific individuals cannot be or are not identified due to the uncontrolled research environment of real-world research. I argued that this problem has far-reaching consequences for our efforts to conduct ethical real-world research since it means we cannot transpose paradigmatic research ethics norms easily outside the proverbial lab. I also argued that this inability to identify can be understood as a matter of low control over a combination of various factors, predominantly the control over the experimental system, the control conditions exerted over various aspects of the environment in which the experiment takes place, and certain control

conditions already presented in that environment. Meeting these norms is not always essential to rendering the research morally permissible. Each norm might have its overriding reasons. However, the identification problem becomes a moral problem when no reasons for overriding outweigh not upholding all the moral norms affected. Consequently, it is imperative to solve the identification problem, especially in the context of real-world research's increasing popularity. This prompts the need for further scholarly reflection on developing custom research ethics guidelines and governance mechanisms or adapting real-world research practice so they can uphold paradigmatic moral norms.

Prologue to Chapter 5

In the previous three chapters, I offered a comprehensive analysis of the ethics of real-world research. That is, I focused on ethical challenges and issues that concern all examples of real-world research: a unifying moral concern in coupling (Chapter 2), a lack of research ethics governance (Chapter 3), and the identification problem making it difficult or impossible to uphold paradigmatic research ethics norms that rely on identifying research subjects (Chapter 4).

In the next two chapters, I zoom in further on the tension between real-world research and upholding paradigmatic research ethics norms by analyzing two distinct case studies of real-world research: real-world AI research (Chapter 5) and live-in laboratories (Chapter 6). In doing so, I aim to accomplish three goals. First, I will show how the issues I have raised so far play out on a concrete case basis. Second, I aim to show that even though, at face value, these examples have stark contrasts, they both present similar ethical concerns due to them being real-world research. Third, these two chapters address specific practitioner audiences, respectively, the generative AI community and the live-in laboratory innovation community, raising awareness about the friction between these real-world research practices and paradigmatic research ethics within these communities.

First, in Chapter 5[36], I provide an analysis of the ethical challenges of real-world research with LLMs and generative AI. Specifically, I ask:

> RQ5: *What challenges arise when we evaluate real-world AI research with paradigmatic research ethics principles?*

I argue that despite its potential epistemic value, real-world AI research faces challenges in meeting moral principles influencing research ethics standards — non-maleficence, beneficence, respect for autonomy, and distributive justice — and that these challenges are exacerbated by absent or imperfect current ethical governance.

---

[36] Published as: Mollen, J. (2025). LLMs beyond the lab: the ethics and epistemics of real-world AI research. *Ethics and Information Technology*, 27(1), 1-11.

5. The Ethics and Epistemics of Real-World AI Research

Abstract

Research under real-world conditions is crucial to the development and deployment of robust AI systems. Exposing large language models to complex use settings yields knowledge about their performance and impact, which cannot be obtained under controlled laboratory conditions or through anticipatory methods. This epistemic need for real-world research is exacerbated by large-language models' opaque internal operations and potential for emergent behavior. However, despite its epistemic value and widespread application, the ethics of real-world AI research has received little scholarly attention. To address this gap, this paper provides an analysis of real-world research with LLMs and generative AI, assessing both its epistemic value and ethical concerns, such as the potential for interpersonal and societal research harms, the increased privatization of AI learning and the unjust distribution of benefits and risks. This paper discusses these concerns alongside four moral principles influencing research ethics standards: non-maleficence, beneficence, respect for autonomy, and distributive justice. I argue that real-world AI research faces challenges in meeting these principles and that these challenges are exacerbated by absent or imperfect current ethical governance. Finally, I chart two distinct but compatible ways forward: through ethical compliance and regulation and through moral education and cultivation.

5.1. Introduction

In March 2023, the Future of Life Institute released an open letter titled 'Pause Giant AI Experiments,' signed by a long list of prominent figures in artificial intelligence research and governance (Future of Life 2023). Prompted by recent developments in the capacities and public deployment of generative AI systems, the letter posited that AI labs were locked in an uncoordinated race to develop and release powerful AI systems into society even though the societal consequences of these technologies were unknowable, uncontrollable, and potentially disastrous. As a solution, the letter urged AI labs to immediately pause the development and training of large language models (LLMs) more powerful than the GPT-4 model for at least six months to understand the systems better, focus on implementing safety protocols for AI design, and develop robust AI governance systems to ensure the safety of powerful AI systems (Future of Life 2023; 2023b). While a pause never materialized, the research and development of large language models has not slowed down since.

This paper focuses on a type of AI research and development that has yet to receive much philosophical attention: research conducted under real-world conditions. Aside from controlled laboratory studies, AI systems are routinely tested in 'the wild' or 'everyday social contexts,' such as their eventual use setting (David & Gross, 2019, p.992). This real-world research is central to developing and deploying robust AI systems. Exposing AI systems to complex and unpredictable socio-technical environments can yield insights about their performance, which cannot be obtained under controlled laboratory conditions. Real-world research can differ from scientific research in that it, for example, does not necessarily employ experimental control techniques such as randomization, control groups, and isolating variables (Ansell & Bartenberger, 2016; 2017) or aim to accept or reject a particular hypothesis as is common in scientific research and experimentation (Popper, 1957; Rheinberger, 1997). Instead, real-world research is broadly more concerned with innovation, group or cluster-level interventions, making the technology work in its use setting and is often characterized by their absence of control to retain the 'natural' representative quality of the research environment (Ansell & Bartenberger, 2016; 2017).

This lack of attention is problematic for at least two reasons. First, as mentioned, real-world AI research is widespread and crucial to AI development and deployment. With the intent to promote innovation, real-world AI research is routinely enabled and encouraged by 'soft law' mechanisms such as 'regulatory sandboxes' (Ranchordas, 2021) or made exempt from many regulatory demands in AI govern-

ance regulations such as the European Union's AI Act (Colonna, 2023). Second, real-world research raises ethical concerns. While there has been increasing scholarly and political attention to the ethics and governance of generative AI systems – such as their capacity to violate copyright laws (Lucchi, 2023), create biased output (Zhou et al., 2024), enable plagiarism (Kwon, 2024) or manipulation (Klenk, 2024) and cause ecological impact (Bender et al., 2021) – similar attention has not been extended to *researching* generative AI systems. However, scholars are increasingly drawing attention to the ethical issues that real-world research with emerging technologies brings about. These issues include the avoidance of democratic accountability by investigators (Taylor, 2021), causing physical harm (Stilgoe, 2020; Colonna, 2023), violating human rights (Amnesty, 2020), the imposition of 'dominating risk' (Maheshwari & Nyholm, 2022), and the unequal ethical demands between various categories of real-world research (Mollen, 2024).

In order to address this gap, this paper provides an analysis of real-world research with generative AI systems and the large language models on which they are built. I will assess both its epistemic value and ethical dimensions. First, I outline the epistemic need for real-world research with large language models. I discuss the limitations of controlled or anticipatory learning methods such as laboratory benchmarking and forecasting and argue that these limitations are exacerbated by large-language models' opaque internal operations and potential for emergent behavior. Second, I argue that this creates an epistemic need to acquire knowledge about large language models through real-world research and outline various potential learning outcomes. Third, I argue that despite its epistemic value, real-world research with AI brings about various ethical concerns that must be taken seriously. I structure these concerns alongside four moral principles that have influenced research ethics standards: non-maleficence, beneficence, respect for autonomy, and (distributive) justice. I then argue that these moral concerns are exacerbated by absent or imperfect current ethical governance. Finally, I discuss two distinct but compatible ways forward regarding embedding research ethics in real-world AI research: through ethical compliance and regulation and through moral education and cultivation.

5.2. The Limits of Controlled and Anticipatory Learning

In this section, I will discuss the limitations of learning methods that allow us to gather knowledge about large language models before they are studied under real-world conditions. I will discuss benchmarking and forecasting. In a nutshell, the

shortcoming of these methods is that they either rely on what can be currently known about the model in a controlled and unrepresentative context or rely on anticipatory or predictive information, which is speculative to a certain degree. Specifically, I argue that these shortcomings are exacerbated by large-language systems' largely opaque internal operations and potential for emergent behavior.

5.2.1. The Limits of Benchmarking

Benchmark tests are standardized software performance tests that measure a system's performance across various tasks and topics (Reuel et al., 2024). Benchmarking allows the evaluation of the quality of the systems or models and the ability to compare this to the performance of other AI systems (Reuel et al., 2024). One example of a language model benchmark is Stanford's Holistic Evaluation of Language Models (HELM) (Liang et al., 2022; Bommasani et al., 2023). HELM involves a multi-metric evaluation of a language model across various scenarios and metrics. These scenarios can involve, for example, answering questions ranging from mathematics to ethics, as well as summarization and information retrieval. Metrics include, among others, fairness, accuracy, bias, robustness, and toxicity (Liang et al., 2022; Bommasani et al., 2023). Benchmark tests can allow for transparent communication to users, regulators, and the larger public about the quality of specific models across various scenarios and metrics and indicate the need to amend the model if low-performance scores are measured.

However, benchmark tests conducted under laboratory conditions face various limitations. First, benchmarks can run into the potential problem of 'restricted scope', in that tests might target only known capabilities and overlook unknown capabilities (Srivastava et al., 2022). It can prove difficult to accurately model the conditions and interactions a large language model might be subject to when embedded in a more extensive socio-technical system (Srivastava et al., 2022). Second, there is a problem of potential 'construct validity': the degree to which a test captures what it aims to assess (Raji et al., 2021; Mökander & Floridi, 2021). For example, particular LLM benchmarks aim to capture normative concepts, such as fairness or safety, yet lack clear philosophical foundations (Mökander & Floridi, 2021). Third, large language models can bring about risks and social consequences – such as the automation of jobs – which cannot be measured at the technology level and thus cannot be captured in a benchmark (Mökander & Floridi, 2021).

5.2.2. The Limits of Forecasting

A form of anticipatory learning about large language models is through various so-called 'foresight' or 'forecasting' approaches (Brey, 2017). These approaches aim, as Brey notes, to

> "project likely, plausible or possible future products, applications, uses and impacts that may result from the further development and introduction of an emerging technology" into a society based on what are inherent or necessary system features or conditions for their realization" (Brey, 2017, p. 179).

One example Brey gives is the Delphi method – a technique that establishes expert consensus on current and potential future developments on a particular issue (Brey, 2017). A recent study employed this method to study the possible impact of large language models on scientific practice (Fecher et al., 2023). Similar anticipatory studies have stressed the social impact of large language models on medical research and care (Clusmann et al., 2023), the labor market (Eloundou et al., 2023), mental health services (Van Heerden et al., 2023), and crime (Europol, 2023), among others.

However, forecasting methods are limited since the complex socio-technical environments that these models aim to operate within make predictions with a high degree of confidence difficult (Van de Poel. There exists disagreement as to the degree to which this shortcoming of forecasting methods can eventually be resolved and, hence, whether the inability to accurately predict the trajectory of a technology is a mere methodological obstacle or an 'ontological' limit (Liebert & Schmidt, 2010). This problem is central to the Collingridge dilemma that states that we have the most control to shape (the trajectory of a) technology when there is little knowledge about its social impact – and vice versa (Collingridge, 1982; Liebert & Schmidt, 2010; Van de Poel, 2016; Kudina & Verbeek, 2019). Additionally, Van de Poel argues that forecasting might focus disproportionately on tantalizing but unlikely scenarios and consequently draw attention away from more realistic but less thought-provoking issues that need attention more (Van de Poel, 2016). In the context of large-language models, we might group existential concerns about machine superintelligence in this corner.

5.2.3. Exacerbating Limitations: System Opaqueness and Emergent Behaviour

To some extent, the shortcomings mentioned above are the case for every technology. However, they do not necessarily apply to *the same degree* for every technology. I

argue that large language models have additional characteristics that make controlled and anticipatory learning more difficult than other technologies: they are 'opaque' technologies and (potentially) capable of 'emergent behavior'. These two features – the opaque nature of large language models and their potential for emergent behavior – further trouble attempts to understand both a system's current and future capacity and behavior.

First, large-language models are 'opaque' technologies. With opacity, I refer to the idea that we have limited access to explanations about an artificial system's inner workings or reasonings (Smith, 2021; Vaassen, 2022). Burrell distinguishes between three sources of 'opacity': either through an (intentional) failure of corporate or state communication, a lack of expertise or technical literacy, or due to the system's inherent features and required scale of use (Burrell, 2016). The latter source is relevant to my point. To take OpenAI's GPT large language model as an example, the number of parameters of GPT-1 grew from about 117 million parameters in 2018 (Hadi et al., 2023) to 1.5 billion (GPT-2) to 175 billion parameters for GPT-3 (Zhang & Li, 2019). Additionally, large language models are trained on massive datasets, making it often difficult to understand the exact makeup of the training data (Bender et al., 2021). The opacity induced by this scale makes fully understanding the current and future behavior of a powerful large-language model difficult.

A second feature of large-language models that might contribute to limited anticipatory and controlled learning is the possibility of 'emergent behavior' (Wei et al., 2022; Hagendorff, 2023; Webb et al., 2023). Emergent behavior refers to the idea that, due to the scale of the models involved and their complex internal interactions, a large language model can produce unpredictable behavior that the system was not necessarily trained for and was absent in smaller versions of the model (Wei et al., 2022). This presents a problem in understanding the capabilities of a larger language model based on the capacities of a smaller version since the scaling could have expanded the capabilities of a model beyond those of the smaller version (Wei et al., 2022). Potential emergent behavior has been observed, however, as Srivastava and colleagues note, "we are unable to reliably predict the scale at which new breakthroughs will happen.." and might "..be unaware of additional breakthroughs that have already occurred but not yet been noticed experimentally" (2022, p, 4). Additionally, Hagendorff claims that traditional benchmark tests cannot detect emergent abilities (2023). Whether this behavior is actually 'emergent,' in the sense that scale causes fundamental changes in the model's behavior, is a current matter of debate. Some have argued that what some label as emergent behavior is better

explained through other means, such as 'metric choices' or 'in-context learning' (Schaeffer et al., 2023; Hodel & West, 2023; Lu et al., 2023). Regardless of the origin of those capacities or what we decide to label as emergent behavior, for my purposes, the point stands that there are difficulties in gaining knowledge about the total range of capacities of large-language models.

So, while controlled and anticipatory methods might teach us how powerful large language models operate under specific controlled conditions, they provide us with little operational understanding and confidence in how the generative AI system might perform under real-world conditions. Here, an epistemic need emerges. In the next section, I discuss the specific learning outcomes that real-world research can offer.

5.3. The Epistemic Value of Real-World AI Research

In this section, I discuss the epistemic value of real-world AI research. Since controlled and anticipatory learning is limited, this creates an epistemic need to acquire knowledge about AI systems through research under real-world conditions. Exposing AI systems to diverse, representative, and unpredictable environments can yield insights about their performance, which are impossible or difficult to obtain under anticipatory laboratory conditions.

First, real-world AI research can show how a particular large language model performs in its potential use setting rather than in a controlled research setting. For example, New York City Public School's AI Policy Lab tested how large language models can aid educational tasks such as lesson planning (GovTech, 2023). The British Department of Education researched whether ChatGPT could aid officials in summarizing and comparing various training plans (Seddon, 2023). Other examples include studies that have explored the impact of LLMs on the development of critical thinking skills in high school classes level (Bitzenbauer, 2023) and their potential to identify errors in student homework and provide them with personalized feedback (Bewersdorff et al., 2023).

Second, real-world research allows the possibility to discover whether a large language model is *comparatively* superior or inferior to another in a specific use context and, thus, which model suits a particular socio-technical environment. For example, the U.S. Department of Defence conducted tests with five different large language models to study to what degree they could improve access times to internal infor-

mation or even help plan responses to potential global conflicts (Manson, 2023). Such tests allow for determining optimal LLMs available in situ.

Third, real-world research allows learning about how generative AI can be successfully embedded within specific institutions. The successful embedding of a novel technology within an organization often goes beyond mere technical capacity but largely depends on social factors. As such, real-world research allows for learning about, for example, which institutional rules and practices can help the effective adoption and use of the AI system or what additions might be necessary to secure responsible embedding, such as digital watermarks to algorithmically identify AI-generated content (Kirchenbauer et al., 2023). Real-world research could thus offer social learning about the successful and responsible embedding of generative AI systems within operations.

Fourth, real-world research allows for monitoring and responding to emergent social impacts of large language models. For example, the EU's Artificial Intelligence Act (AIA) mandates that the providers of high-risk AI systems must engage with 'post-marketing surveillance' to monitor, document, and analyze the performance of these systems throughout their life cycle (Mökander et al., 2022). Post-market surveillance refers to a set of monitoring activities a manufacturer has to perform to ensure the performance and safety of their product *after* it has been released on the market (Pane et al., 2019; Beckers et al., 2021). During post-marketing surveillance, providers are expected to report serious malfunctions and take immediate action to either correct this malfunction or withdraw it from the market (Mökander et al., 2022). Through these measures, the performance and continued safety of these products can be closely monitored and, ideally, withdrawn from the market in the case of negative social consequences.

Fifth, real-world research provides an opportunity to learn about a generative AI system's normative and moral consequences, for example by offering the chance to test and observe whether a system meets particular ethical requirements in-situ. The Dutch government's Impact Assessment Fundamental Rights and Algorithms notes that real-world test beds can help identify harms to fundamental rights before such models are publicly released (Janssen, 2020; Ministerie van Algemene Zaken, 2022). Harbers and Overdiek have also argued that real-world living labs could contribute to ethical AI design, development, and deployment (Harbers & Overdiek, 2022). Mökander and colleagues have recently proposed 'ethics-based auditing,' which assesses large language models to determine their consistency with relevant moral values (Mökander & Floridi, 2021).

Finally, Van de Poel has argued that since research environments are 'small scale' compared to learning from a more public-wide market release, potential negative consequences will be comparatively minor (Van de Poel, 2017). Van de Poel refers to this as 'learning-by-experimentation', as opposed to learning-by-anticipation or learning-by-doing (Van de Poel, 2017). Since research environments are (ideally) closely monitored, costs will be 'limited' since these costs happen on a small scale (Van de Poel, 2017). As such, we will know at an early stage when negative consequences arise and they are more easily mendable.

Thus, real-world AI research meets a critical epistemic need since it can provide valuable insights into the successful development and embedding of generative AI systems that we cannot acquire through controlled or anticipatory methods. However, this learning also raises ethical concerns, which I will outline in the next section.

5.4. The Ethics of Real-World AI Research

In this section, I discuss various ethical dimensions of real-world research with generative AI and large language models. Despite its epistemic value, real-world AI research also raises ethical concerns. I organize these ethical concerns along four moral principles that underpin many legal, professional, and moral standards regarding ethical research: 'non-maleficence,' 'beneficence,' 'justice,' and 'respect for personal autonomy' (Beauchamp & Childress, 1994). These principles were outlined by Beauchamp and Childress as being central to ethical conduct in healthcare and the biomedical and behavioral sciences. These principles have since been influential in shaping much of contemporary research ethics guidelines and AI ethics guidelines such as the EU's Ethics Guidelines for Trustworthy Artificial Intelligence or OECD's Recommendation of the Council on Artificial Intelligence (Nikolinakos, 2023; Porter et al., 2024). Additionally, they also provided the basis for Van de Poel's ethical framework for the moral evaluation of introducing experimental technology into society (Van de Poel, 2016). As such, I consider them an apt starting point through which to analyze the research ethics of real-world AI research. I do not aim to defend or criticize this framework or particular interpretations of the moral principles involved or argue that this list intends to be complete. Instead, I use these moral principles to capture and organize a wide range of relevant ethical issues in real-world AI research and discuss how real-world AI research might bring about context-specific challenges in addressing these issues.

5.4.1. Non-maleficence

First, the moral principle of non-maleficence refers to the idea that research interventions should 'do no harm' (Beauchamp & Childress, 1994). Here, I define harm not merely in a physical sense but, like Feinberg, as any 'wrongful setback' to or 'thwarting' of an interest, such as the violation of a right (Feinberg, 1984). Researchers should not cause harm or should prevent harm from arising as a consequence of the research intervention.

The risks that real-world AI research might bring about can vary. For example, Colonna has argued that testing artificial intelligence under real-world conditions can present "risks to individual's health, safety, and fundamental rights, as well as broader societal concerns" (Colonna, 2023, p.28). An example of such a broader societal concern is the environmental impact of AI systems. Due to the energy consumption and global resources required during the entire lifespan of an AI system, scholars have increasingly drawn attention to the carbon cost and environmental impact of AI systems (Dhar, 2020; Bender et al., 2021). Hence, the (real-world) research and development of powerful large language models - given their current energy consumption - will further impact the environment, increase the carbon footprint, and contribute negatively towards mitigating climate change (Dobbe & Whittaker, 2019; McDonald et al., 2022; Lakim et al., 2022; Rillig et al., 2023). This means that even if the potential negative consequences in real-world research settings will be comparatively more minor, generative AI systems or large language models can still carry risks, some of which may be substantial. When researching these systems on a group level, we effectively expose populations interacting with these systems to these risks.

Real-world AI research, however, poses challenges to prevent or mitigate harm for at least two reasons. First, when research is conducted within a real-world environment, predicting, containing, and identifying risks - or even identifying which persons might be affected by the intervention can become more difficult due to the interconnected and complex real-world environments in which some AI systems are tested, as I argued in Chapter 4. If researchers cannot identify who is harmed during or after the experiment, compensating or remedying harm becomes difficult or impossible.

Second, it is unclear how early detection of negative consequences might lead to adjustments to the design or implementation of language learning models. One of the possible benefits of learning about technology through closely monitored small-scale introduction is that, ideally speaking, knowledge about negative consequences

can be quickly fed back into improving either the design or embedding process. This idea of controlled, iterative learning also underpins much of post-marketing monitoring and regulatory sandboxes. However, large language models are complex digital technologies. Unlike physical devices, such as toasters or cars, that can be redesigned in response to specific safety concerns, large language models are complex, adaptive systems that do not allow for straightforward design modifications in response to individual adverse outcomes.[37] At best, monitoring might prompt a recall of a particular technology. In some cases – see some of the examples in Section 3 – parties testing out a particular large language model only have (paid) access to use the model and are not able to make changes to the underlying model when negative consequences might arise. Instead, they only have the power to decide how they will use the model or whether they will use it at all. Hence, even if negative consequences arise in a real-world test, this does not necessarily mean that these insights will be translated back into fundamental changes to the models.

5.4.2. Beneficence

Second, the moral principle of beneficence prescribes that aside from avoiding harm, researchers are also obligated to "act to the benefit of others" (Beauchamp & Childress, 1994, p. 203). According to Van de Poel, this includes, for example, "obligations to take away existing harm, or to prevent harm or risks that do not originate in the experiment, to produce more good than harm, to create or increase benefits" (Van de Poel, 2016, p. 676). If real-world AI research brings about risks or harms to particular persons or groups, such research should at the very least be conducted with the intention – and under the reasonable belief – that the research with the generative AI system will bring social value into the world, either by directly benefitting people's lives or, for example, by lowering the demands on public resources through more efficient operations.

One potential challenge to this aim of beneficent real-world AI research is the increased privatization of AI research. In recent years, the center of gravity of AI research and development has increasingly shifted away from (public) academic institutions to private companies (Jurowetzki et al., 2021; Gizinski et al., 2024). As the *who* of AI research transitions towards industry, this changes *what* is being learned and *who* has the power to decide to what is being learned. The AI industry plays a

[37] I want to thank an anonymous reviewer for stressing this point.

large role in identifying, influencing, and shaping the 'problems' that receive research focus and funding (Khanal et al., 2024). Consequently, private interests can constrain research scope or funding and limit research topics not in line with the corporate interest but which might be socially relevant (Jurowetzki et al., 2021). This way, corporate interests set the AI research agenda, which might not necessarily align with societal goals. For example, industry-driven learning might favor short-term monetization and competitive advantages and hold lower expectations for the social value of their research or other considerations such as environmental costs, societal externalities, and ethical challenges (Bender et al., 2021; Jurowetzki et al., 2021). This, Jurowetzki and colleagues argue, "bolsters the case for increasing AI research capabilities in academia and government in order to ensure that public interests can continue playing an active role in monitoring and shaping the trajectory of powerful AI systems" (2021, p. 2).

5.4.3. Respect for Autonomy

Respect for autonomy refers to the obligations of researchers to respect the autonomy of persons or groups involved in the research (Van de Poel, 2016). Beauchamp and Childress hold that "to respect autonomous agents is to acknowledge their right to hold views, to make choices, and to take actions based on their values and beliefs" (Beauchamp & Childress, 1994, p. 106). Persons have a right to make autonomous decisions in that they should have control over their own lives, bodies, and data and make decisions about them according to their reasons, motives, and interests. Since research can intervene with a person's autonomy, particular research ethics mechanisms, such as informed consent and withdrawal procedures, aim to help safeguard a person's autonomy (Van de Poel, 2016).

Here, real-world AI research raises various ethical concerns. One concern is the question of availability and access to information about the research. Since real-world AI research takes place in 'natural' environments, people might not be aware that they are part of a research project without being adequately informed. If people are unaware that they are part of a research project, they cannot make an informed decision to participate in the research and thereby consent to its potential associated risks and benefits. However, even if a person is aware of the research happening, issues arise regarding the ability to opt-out. For example, how can a person meaningfully opt-out from interacting with a generative AI system that is tested in an area that is difficult or costly to avoid, such as a place of work or government institutions? Additionally, there are concerns regarding data ownership. How can subjects

exposed to real-world AI research keep control over their data (mainly when industry parties might conduct such research), and what rights and abilities do they have to amend or withdraw their data after the fact?

5.4.4. Distributive Justice

The principle of distributive justice in research ethics refers to researchers' obligations regarding a just distribution of the research's benefits and risks (Beauchamp & Childress, 1994). This includes norms such as fair subject selection, protecting 'vulnerable subjects', or ensuring that research in which 'vulnerable subjects' participate is beneficial to them (Van de Poel, 2016).

Real-world AI research can bring about various issues of distributive justice. Due to a lack of ethical governance (I will expand on this in the next section), there may be a tendency to conduct research in areas or regions with less regulatory oversight or among individuals or groups who lack sufficient awareness of these risks. This would mean that risks are disproportionately placed on those communities that enjoy the least protection. Additionally, it might be difficult to provide safeguards for vulnerable persons or groups when these persons or groups are challenging to identify in the real world. If researchers are not aware of the exact demographic makeup of their subject pool, it will be difficult to exclude – or award additional protections - vulnerable individual subjects or groups.

Another question concerns how affected people and groups can share in the benefits of real-world AI research that is subjecting them to particular risks. Here, the issue of increased AI privatization also plays a role in *who* benefits from this learning. As mentioned, knowledge about AI systems or their performance is increasingly concentrated within private companies. This data could be difficult or undesirable to share with academia for proprietary reasons or to retain a market advantage (Jurowetzki et al., 2021) and thus difficult to reproduce and replicate (Gizinski et al., 2024) or made subject to independent ethical scrutiny (Resseguier & Ufert, 2024). Even when public institutions run their own tests with embedding, particularly instances of generative AI, such experiments can still benefit corporate interest if public institutions use systems developed by industry and firmly embed them within their operations, potentially leading to a lock-in problem.

5.5. A Lack of Research Ethics Governance of Real-World AI Research

So far, I have described a tension between the epistemic value of real-world AI research and the various ethical concerns this type of research can bring about. This prompts questions regarding the need for external scrutiny. However, as I argued already in Chapter 3, real-world research can be characterized by a lack of research ethics governance. In this section, I discuss this lack in the concrete case of real-world AI research. In doing so, I show how this limited research ethics governance brings about difficulty in navigating a tension between epistemic value and ethical concerns within the field of real-world AI research and exacerbates the problems mentioned in the previous section in failing to uphold moral principles.

Generally, research ethics governance mechanisms – such as guidelines, protocols, or ethical review boards or committees – aim to address or (help) navigate the ethical tensions described in the previous section. They can do so by providing action-guiding norms or through various research ethics mechanisms that provide a means of reviewing research proposals (and their research's risks and risk mitigation strategies) and holding researchers accountable for research malpractice and subject redress.

However, research under real-world conditions with generative AI lacks research ethics governance (Mollen, 2024). While clear research ethics regulatory demands generally bind scientific research, such demands are often absent in research conducted by industry or public parties. While there has been an increase in AI guidelines and ethics codes, Munn has argued that these ethical principles are largely useless and do not impact practice since they are 'meaningless' (contested or incoherent), 'isolated' (applied to domains that ignore ethics), and 'toothless' (without consequence or in-line with industry interest) (Munn, 2023, p. 872). This leaves people and groups vulnerable since there are no mechanisms for external scrutiny, and people are not effectively given control to counter-act experimental impositions.

The absence of research ethics governance can also enable the evasion of ethical demands elsewhere. As I argued earlier in Chapter 4 when different research ethics regulatory demands are placed on two research domains, one research domain can avoid such regulatory demands by placing particular research activities outside the scope of the regulatory demands they are subject to (Metcalf & Crawford, 2016; Colonna, 2023; Mollen, 2024). Since the private AI developers that engage with real-world AI research are not bound by the same regulatory demands as scientific AI researchers, this allows for the evasion of these regulatory demands by the latter through private-public collaboration. For example, while specific data might not be

captured without the subjects' consent by scientific research, when no such demands are placed on corporate researchers, the latter could collect this data. At that point, it becomes public data that can be used. In this way, the absence of research ethics governance in one domain can come at the expense of those whom other ethical demands aim to protect.

Even if such research is conducted by parties operating within an ethically regulated domain – for example, scientific publicly-funded research – the available ethical guidelines or protocols might not help address researchers' moral and regulatory challenges. For example, AI and data scholars have increasingly called for research ethics reforms to address current limitations (Vitak et al., 2017; Raymond, 2019). Resseguier and Ufert, for example, have argued in favor of adaptations of current research ethics standards and mechanisms to better asses scientific AI research, such as submitting an assessment of risks and harms to communities, society at large, and the environment when the AI is deployed 'in the real world', extending the period of when risks and harms are considered from the research stage to when the AI systems are deployed (2023). Additionally, Resseguier and Ufert argue that much of the data that fuels current AI research comes from scraping existing data, using existing data sets, or collecting data "in the wild" (2023, p. 147). Under current research ethical guidelines, this data is often considered exempt from ethical review (Ada Lovelace Institute, 2022). While this data might be innocuous in the original study, it can be re-combined to create more problematic datasets (Metcalf & Crawford, 2016). Hence, adapted research ethics for AI research needs to be sensitive to this kind of data collection.

However, as long as research under real-world conditions is conducted solely or in partnership with parties not bound by these ethical demands – such as many industry parties – these research ethics reforms only target scientific AI research at best. Real-world research with artificial intelligence often involves research collaborations between private, public, and knowledge institutions (Ada Lovelace Institute, 2022). Such collaborations between stakeholders can cause confusion about how (moral) responsibilities should be divided and how particular ethical concerns can be navigated or resolved in the case of conflicting values or interests within the research consortium.

5.6. Moving Forward: Embedding Ethics within Real-World AI Research

The above section presents a persuasive case to ameliorate the current situation in which much of real-world AI research is conducted under imperfect ethical governance – or in its complete absence. In this section, I will briefly discuss the benefits and drawbacks of two distinct but mutually compatible approaches to embedding research ethics within real-world AI research: through ethical compliance and regulation and through moral education and cultivation.

The first approach relies on ethical governance through regulation, such as mandatory ethical compliance or an institutional review board review. One example of this approach is the conditions the EU's AI Act places on research with high-risk AI outside the scope of regulatory sandboxes. These include requiring informed consent, additional protections for vulnerable populations, the protection of personal data, removing personal data after persons have withdrawn their consent, outlining the roles and responsibilities of all parties involved, and creating a real-world testing plan detailing the goals and duration of the research which needs to be registered in an EU-wide database and submitted to 'competent market surveillance authorities' (AI Act 72b).

A benefit of such ethical governance is that it is mandatory, creating a concrete incentive for industry and public parties aiming to research a particular AI system under real-world conditions. Additionally, it provides governments with a 'checkpoint' to assess and influence what kind of research is conducted with (generative) AI under real-world conditions, ensuring that the research creates public value. On the other hand, mandatory regulations can also bring about a 'checklist' ethics mentality, creating additional costs and demands for government oversight agencies and practical and conceptual challenges to meeting these demands when conducting research under real-world conditions, for example, difficulties in obtaining informed consent or protecting vulnerable groups when it is difficult to identify research subjects.

An alternative approach aims to foster ethical AI research through non-mandatory incentive structures such as independent review boards providing research ethics advice (Polonioli et al., 2023), workshops, design activities, games, or roleplaying for practitioners to create increased awareness about the moral dimensions of their research practices (Wong et al., 2020), ethics training (Hagendorff, 2022), conference and journal standards (Polonioli et al., 2023), etc. One example would be the Dutch Fundamental Rights and Algorithm Impact Assessment

(FRAIA), a human rights dialogue and reflection tool for developers or deployers of algorithmic systems.

A benefit of this approach is that it aims to motivate, interest, and cultivate a researcher's conviction to do good rather than to be merely compliant with mandatory regulation. However, an apparent shortcoming of this approach is its largely self-regulating nature, meaning that if these approaches are unsuccessful or purposefully neglected, they leave little protection for those affected by the research intervention.

5.7. Conclusion

In this paper, I have discussed the epistemic value of real-world AI research and the ethical concerns this type of research brings about. While generative AI and large language models hold great promise, they must be developed in a manner that is ethical and consistent with moral principles. While there is a clear epistemic need for real-world AI research – exacerbated by large-language models' opaque internal operations and potential for emergent behavior – this does not mean this research should be conducted without the ethical guardrails we find in other types of (scientific) research. Currently, real-world AI research is conducted in a space that lacks proper ethical governance, leaving persons and groups without due protection and exacerbating real-world AI research's moral concerns. Hence, we should strive to ameliorate the current situation by drawing from two distinct but mutually compatible approaches to embedding research ethics within real-world AI research: ethical compliance and regulation and moral education and cultivation. While these methods might have their respective downsides, a balanced approach incorporating ethics in real-world AI research is not only necessary but overdue.

Prologue to Chapter 6

To reiterate, in Chapters 5 and 6, I show how the themes of this dissertation so far play out on a concrete case basis. I aim to show that even though, at face value, these examples have stark contrasts, they both present similar ethical concerns due to them being real-world research. In doing so, I address specific practitioner audiences, respectively, the generative AI community and the live-in laboratory community, raising awareness about the friction between these real-world research practices and paradigmatic research ethics within these communities.

After analyzing real-world AI research in Chapter 5, I continue my focus in Chapter 6[38] on the ethics of real-world research on a case-study basis by turning to live-in laboratories, which are homes that are built as experimental living environments to test the performance of novel technologies on their residents.

A major theme in this thesis (and especially Chapter 2) is that real-world research 'couples' itself with daily life. This theme is perhaps most pronounced in live-in laboratories, which 'couple' in a clear sense daily life with research participation. Residents live in an experiment, and to no longer be research participants, they would have to move house. A second theme is that paradigmatic research ethics norms are difficult to uphold when applied to real-world research.

I show how these two themes come together in live-in laboratories by analyzing how the 'right to withdraw' — a research ethics norm that grants research subjects the ability to withdraw from research without penalty or coercive influences in order to safeguard the voluntary status of research participation – conflicts with this real-world research practice. Specifically, I ask:

RQ6: *How, if at all, does the right to withdraw conflict with live-in laboratory research?*

The chapter argues that live-in laboratory research conflicts with the right to withdraw, and if we were to take the right to withdraw seriously, then the practice of coupling a participant's main residence to research participation would be ethically problematic.

---

[38] Published as: Mollen, J. (2023). Moving out of the Human Vivarium: Live-in Laboratories and the Right to Withdraw. Journal of Ethics and Emerging Technologies, 33(1), 1.

6. The Human Vivarium: Live-in Laboratories and the Right to Withdraw

Abstract

Live-in laboratories are homes constructed to be research environments in which the performance of novel technologies can be tested in real-world settings on its residents. When people's homes are turned into the site of experiments, the inhabitants become research subjects. This paper argues that when live-in laboratories function as a participant's main residence, they constrain an individual's so-called 'right to withdraw.' The right to withdraw is a paradigmatic research ethics norm. However, withdrawing from the live-in laboratory as a participant's main residence means losing one's home, which creates negative financial and psychological consequences for participants. I will argue that such costs conflict with a participant's right to withdraw on two counts. First, the exit costs from the live-in laboratory constitute a penalty, and second, the costs of withdrawing from the live-in laboratory function as a constraint on a participant's liberty. The paper concludes that (i) the right to withdraw is a necessary condition for the ethical permissibility of modern live-in lab experiments and concludes (ii) the practice of making an experimental home a participant's main residence is ethically problematic.

6.1. Introduction

What if withdrawing from an experiment means losing your home (Taylor, 2020)? In the last two decades, living environments have been constructed for the explicit purpose of performance and hypothesis testing while hosting participants as residents, such as the MIT PlaceLab or Georgia Tech's Aware Home. These experimental living environments, often called live-in laboratories, aim to bridge the research benefits of a controlled laboratory setting with extensive fieldwork (Intille et al., 2005).

But, when homes become laboratories, their inhabitants become research participants. Live-in laboratories exemplify an intimate relationship with their research participants that few research methodologies possess. Residents are subjected to a perpetual state of exposure to a variety of experimental interventions and forms of data capture depending on the technologies tested. Live-in laboratories thus have strong research ethical implications. Regardless, while individual studies in live-in laboratories might be subject to ethical review, the ethics of live-in laboratories as a research platform have received limited academic scrutiny.

Research ethics, as a matter of governance, is upheld within the scope of an institution's authority or legal jurisdiction. This means researchers or institutions external or in collaboration with a university do not fall under the scope of its ethical review and are not bound to their ethical guidelines. Live-in laboratories often operate as locations for collaboration between knowledge institutes and public and private parties. Here, researchers with various backgrounds can research, test, and develop new solutions or technologies in a near-to-real-use setting. Consequently, while university scientists might conduct research with or on live-in laboratory residents, they might be required to meet research ethics requirements as outlined by their institution. However, such guidelines do not necessarily apply to non-academic researchers.

Urban environments, both public and private, are increasingly framed as experimental locations where solutions for societal challenges can be found through research and technological innovation (Maas et al., 2017; Baccarne et al., 2014). While experimental practices outside the laboratory are bound by positive law, what is missing, as Taylor notes, is an "interrogation of urban experimentation that takes seriously the issue of research on human subjects, and asks what norms, rules and boundaries are appropriate" (Taylor 2020, 1903). This paper provides such an interrogation.

Taylor has suggested framing urban technological experimentation through a research ethics lens (Taylor, 2020). One of the practical features of research ethics is that it awards research participants what Taylor calls 'avenues of resistance' against asymmetrical power relations between researcher and participants (Taylor, 2020). Such 'avenues of resistance,' for example, preserve participants' freedom from constraints that urge a certain action and provide participants with a certain level of control over potential research risks that they are subject to but do not necessarily control or benefit from. This lens is applicable to live-in laboratories since they are an experimental apparatus that functions as a research methodology to conduct hypothesis and performance testing on and with human subjects. However, it is exactly such resistance that the live-in laboratory renders ineffective.

In those cases where a live-in laboratory functions as a research participant's main residence, withdrawing causes negative consequences for participants, constraining their liberty to exercise their right to withdraw from research effectively. This right is an ever-present ethical principle in contemporary moral codes regulating research on human participants and functions as an important mechanism that helps realize the bioethical principle of autonomy in the conduct of an experiment.

The paper proceeds as follows. First, I define live-in laboratories and explain how withdrawing poses negative consequences for participants. Next, I describe the residents of live-in laboratories as a subject pool that has received limited research ethical attention. I then argue that if an experiment is ethically permissible, a participant is free to exercise their right to withdraw freely without penalty. Then, I show that the cost of withdrawing from a live-in laboratory qualifies as a penalty and that the (unintended) threat of said costs acts as an unjust controlling influence on a participant's liberty to exercise their right to withdraw. Finally, I conclude that the live-in laboratory is an ethically problematic experimental setup and suggest that investigators should aim to nullify the associated costs of withdrawal or only conduct research on temporary residents who do not face exit costs.

6.2. Experimentation and the Live-in Laboratory

Live-in laboratories are experimental homes that are used to either study how persons interact with a certain technology, study persons within an instrumented domestic environment, or test the performance of a technology in a real living environment inhabited by humans. Live-in laboratories vary in scope, scale, and focus. What binds them is that they are real living environments created for the

purpose of hypothesis and performance testing. They are often real homes with residents constructed for research purposes.[39]

This paper focuses on two types of domestic live-in laboratories: 'Visited Places' and 'Lived-in Places', first identified by Alavi and colleagues (2020). The main difference between these two types of live-in laboratories is the duration of occupancy. Visited Places are live-in laboratories that host participants for a few hours or days per week and thus are temporary places of living. (Alavi et al., 2020). Lived-in Places host participants, or residents, for several years and function as a participant's main residence (Alavi et al., 2020).

Let's turn to two examples to clarify the difference. In 2004, MIT constructed the PlaceLab, a live-in laboratory to study domestic ubiquitous technologies (Intille et al., 2005). The PlaceLab is a 1000 sq. ft. apartment embedded with a myriad of sensors, including light and infrared cameras, environmental sensors, microphones and motion sensors. As participants live in PlaceLab for a few weeks at most, this is a 'Visited Place'.

Contrast this with the 'DreamHus' (Frisian for 'dream house'), which are part of the Delft University Technology campus (Dreamhus, 2021). Standing on the site of 'The Green Village', a real-world testbed for sustainable technologies, the 'DreamHus' are built in the image of three 1970s Dutch row houses with the aim to test potential innovative solutions to make more sustainable housing. The aim is to scale up efficient solutions to the general (Dutch) housing stock. Current experiments done by an assembly of researchers, students, and innovators within the Dreamhus include "solutions in the field of energy, healthy indoor climate, water, heating, insulation, ICT, IoT, and Smart homes" (Dreamhus, 2021). These three houses are inhabited for two years maximum, with additional studios for students who can stay for up to five years. These live-in laboratories are Lived-in Places since they function as an occupant's main residence.

This paper is especially concerned with live-in laboratories being used as Lived-in Places (LIP from now on). It is unclear exactly how widespread this phenomenon is. However, a recent study looking at the living lab literature between 1999 and 2018 found 19 instances mentioned in the literature sampled (Alavi et al., 2020). Furthermore, there are plans for an entire live-in lab neighborhood called

[39] There are also examples of offices being designed and built as live-in laboratories, such as the Smart Living Lab in Switzerland (Alavi et al., 2020), which feature experimental and digital technologies that put the space in a constant experimental state.

Brandevoort 2, which aims to construct a complete, digitally connected neighborhood in which residents can be continuous research participants and sell their data for rent reductions (for a more thorough discussion of the Brandevoort 2 project, see (Taylor, 2020)). LIPs are a research methodology using the live-in laboratory as an experimental apparatus and its inhabitants as research subjects. Regardless, such live-in laboratories have received barely any ethical scrutiny to date (Taylor, 2020). With such practices happening right now – and with more similar projects in the pipeline – an exploration of the research ethics of live-in laboratories is timely and necessary.

Specifically, I will analysis LIP's through the perspective of the right to withdraw without penalty, which is a paradigmatic norm in contemporary scientific moral codes regulating research on human participants. The right protects the participant's ability to withdraw their consent to participate in a research experiment or trial at any time and, by effect, stop their participation in said experiment or trial without 'retribution', 'reprisal', 'penalty', or 'loss of benefits' (Schaefer & Wertheimer, 2010; Edwards, 2005; Holm, 2011). However, when would extent this norm to participants in a LIP, this paper observes a friction: the costs of withdrawing from an LIP seem to conflict with a participant's right to withdraw from research without penalty (abbreviated as RTW onwards).

6.3. The Consequences of Withdrawing

Consider the following. A team of researchers has developed a technology (let's call it "T") and would like to gather data on people's interaction with T in a domestic setting. Participant Petra gives informed consent to have T tested and monitored in their home. The research team comes to Petra's home, installs T, and takes their leave. During the experiment, Petra, for whatever reason, changes her mind. No longer wishing to partake as a participant in the experiment, Petra informs the researchers and withdraws their informed consent to participate. The research team removes T, leaving Petra and her house as before the experiment. In this scenario, when a participant withdraws from an experiment in a home that temporarily had become an experimental site, the home returns to its state before it is instrumented. Withdrawing from the experiment came at no cost.

Let us compare this to what withdrawing from an LIP would look like. We take the same team of researchers who have developed technology T and want to gather data on people's interaction with T in a domestic setting. Instead of introducing T

within an existing domestic setting of Petra, they decide to construct their own domestic setting – a live-in lab that acts as a home. Again, they invite Petra, who gives their informed consent to participate in the experiment, live in the live-in lab, and have their interactions with T monitored. Later, Petra changes their mind about their research participation and informs the researchers that they will be withdrawing their consent to participate in the experiment. Now, what happens? As we saw, instead of introducing T to a home, Petra is introduced to the home. Since Petra is the addition to the LIP and not vice versa, we will remove Petra from the LIP.

This is an important difference. While the removal of T from a traditional research setting comes at no cost for Petra, removing Petra from the live-in laboratory comes at a significant cost for her. This consequence is the same if Petra either withdraws themselves from the home or if investigators remove Petra from the home: they are removed from their home and daily life and have to move.

This leaves a participant in an undesirable situation where their housing is contingent on their research participation. If there is no 'baseline home' to return to, which is the case since the LIP is a participant's main residence, then these participants need to find a new house. Moving house, also known as residential mobility, has several costly implications for a participant, which I will now outline.

First, moving house inflicts economic costs upon participants who wish to withdraw. While the financial cost naturally varies based on location, moving is never free. Deposit, mortgage costs, broker fees, estate agent fees, insurance, legal fees, postal redirection, removal, and moving companies are but a few examples of the types of financial costs that moving can inflict. As an indication, according to the UK's non-profit Consumers' Association, the average cost of moving house in 2020 was around £7000,- (Maunder, 2020). Another point to make is that live-in laboratories might offer residencies below market rent, increasing rent costs for those who (have) to move back to the non-instrumented housing stock.

Secondly, residential mobility has an impact on a person's mental health. Research suggests that there is a link between residential mobility and poorer mental health (Morris et al., 2017). This link seems strongest for adolescents and children. Morris and colleagues outline several pathways through which this effect operates, including weakened social ties, disturbance of social networks, social stress, household disruption, and social isolation (Morris et al., 2017).

Additionally, these costs do not happen in a vacuum. The above-mentioned costs are, in fact, aggravated by their socio-economic context, which, while not a cost in itself, impact the capacity of a participant to move house successfully. For example,

there needs to be available housing to begin with. This greatly depends on local housing situations. Available housing also needs to be affordable to the participant who withdraws from the LIP. Hence, research participants from lower socio-economic classes would have a harder time finding replacement housing, considering factors such as long waiting lists for government-sponsored social housing, a disconnect between increasing rent prices in urban areas, and increased wages and minimum income requirements for rental homes. This is a problem since a live-in laboratory would likely attract "experimental subjects who are already on the receiving end of power asymmetries" (Taylor 2020, 1908). Those who will be willing to live in the LIP or feel drawn to its potential lower market rent will be from financially more vulnerable demographics: students, renters, those who qualify for social housing, etc.

Furthermore, moving house is never immediate. This raises questions about the participant's immediate housing status. If there is no immediate alternative, a research participation termination amounts to putting a former participant on the street. If participants are allowed to continue living in the LIP after research participation has ended for a certain period, questions arise concerning the experimental technology present in the LIP while the resident is no longer a research participant. Will these remain operational, but will collected data be stopped or destroyed? Will these technologies be removed or turned off? When the LIP is part of an experimental neighborhood of live-in laboratories, can we credibly say the person has withdrawn from the experiment if all their neighbors, the neighborhood, or the immediate area surrounding their home is still the subject of research? Such unresolved questions might leave an ex-participant in a state of undesirable uncertainty.

An additional problem can emerge when it is not a single individual who inhabits a live-in laboratory but instead a group or family.[40] Cohabitation is a very common living arrangement. Unless clear regulations were in place to prevent it, it is highly plausible that families or other forms of cohabitation might take residence in a live-in laboratory if it is suited to host more than one resident. In fact, couples do live in the Dreamhûses, the live-in laboratories part of Delft University of Technology's Green Village (2023). This presents an interesting problem for live-in laboratories in particular and any form of research that deals with collective forms of subject participation: what if a member of the cohabiting unit wants to withdraw from the experiment and (the) other(s) do not?

---

[40] Thanks to an anonymous reviewer for suggesting this issue.

I will distinguish three distinct scenarios that might follow such a predicament. First, if the subject/resident in question decides to withdraw and move out, they face the same constraints as laid out in this chapter so far. Secondly, it might be possible that the resident who wishes to withdraw does not wish to move out or, at least, is not able to move out right away. This might be due to the nature of the relationship of the residents. For example, a couple might reasonably want to keep living together, and parents cannot abandon their children. Alternatively, it might be due to the above-mentioned constraints, such as market forces and the ability of the resident to afford to move. Regardless, in such a scenario, we are presented again with the problem presented in the previous section: where a participant cannot move out immediately and is potentially, by proxy, still involved in an experiment because their neighbors are. In this case, this would be their cohabitants, and the challenge of successful withdrawal seems even more pronounced. Thirdly, it might be the case that due to one person wishing to withdraw, everyone else either has - or feels too obliged - to withdraw too. This collective withdrawal can either be imposed from an organization running a live-in lab or from individuals valuing their partnership, family, friendship, etc., above their participation and residence. Additionally, children, for example, might have little agency regarding their withdrawal. They can't stay or withdraw without their parents or guardians. The research ethics of underage residents/subjects is the additional problem of live-in laboratories that I do not have the space to address in this paper.

These group dynamics pose additional controlling influences on a resident's decision to withdraw. Co-habiting a live-in laboratory with a partner, family, or friends could very well influence a participant's decision to withdraw since if they chose to do so, either they would have to leave their co-habitation unit or the whole unit would leave the experiment. I expand on controlling influences in section 6.

Finally, it is important to note that live-in laboratories might have specific conditions under which persons are able to inhabit them. These conditions can influence the degree to which withdrawing causes certain consequences. The exact site-specific conditions of a given live-in laboratory are outside the scope of this paper. However, for the sake of providing an example to this point, I will briefly outline several conditions concerning the aforementioned live-in laboratories based on the earlier typology (Visited Places vs Lived-in Places).

First, considering the two LIPs that were mentioned earlier, Green Village and KTH Live-in Lab, residents receive a rental contract for housing that is equal to or below market rent[41][42]. Contracts are offered for a set period, ranging from one year to a maximum of a few years. In return, through living their daily life and interacting with a variety of experimental systems, certain technologies can be tested and developed. Turning to the VPs, MIT PlaceLab residents were reportedly volunteers (Roberts, 2011). While not mentioned, I take this to mean they did not pay rent and potentially received (limited) benefits. This would be plausible given that residents only stay in the Placelab for up to a week or two.

A recurring selection criterion for residents seems to be their interest in the experimental work conducted at the live-in laboratory (KTH 2020). Participants/residents are partially selected based on their motivation and personal connection to overall research themes. It is plausible that this will translate into a more interested, engaged, and complacent resident body, increasing the likelihood of a smooth relationship with residents during their stay. Having an altruistic sense, one has the opportunity to contribute to problems on research themes that they value – say sustainability – and runs the possibility of not only keeping participants engaged but also morally bound to the project.

6.4. Residents as Unrecognized Human Subjects

Residents of live-in laboratories are a human subject pool essential to the live-in laboratory as a research methodology that aims to emulate a near-to-real-use setting. However, LIPs are not being classified as human experimentation due to two main reasons. First, the scope of research ethical regulation is strongly tied to those institutions that apply for federal or governmental funding, leaving the live-in laboratories of private parties outside this scope. Additionally, when live-in laboratories are part of a collaboration between knowledge institutions and public and private parties (so-called triple helix collaborations), research ethical obligations might get obfuscated. Secondly, while individual studies conducted in LIPs might

41 In the case of the Green Village, personal correspondence with staff informed me that their housing rent is below market value. The housing stock consists of studio's (generally for students) and larger family homes, housed by individuals or couples.

42 An application post for residency in the KTH Live-in Lab notes how rent will mirror other apartments the university offers (KTH 2020). The price quoted is 5000-6000 SEK/month. The housing stock consists of shared (student) housing.

meet the criteria for counting as research (with human subjects), the LIP and the act of living in a LIP itself are not research. Instead, they constitute the creation of a continuously available and exposed subject pool. Residents are exposed to a variety of research practices that may or may not qualify as human subject research, yet the LIP itself remains outside of regulatory scope. I will expand on these points below.

Research ethics regulation is commonly applied alongside the institutional boundaries of universities or similar research institutions. In order to qualify for or attract governmental funding, universities, e.d. have to comply with the funding organizations' ethical review regulations (Moffat, 2010). For example, the US Federal Policy for the Protection of Human Subjects, also known as the Common Rule, only applies to behavioral and biomedical research that receives federal US funding and is conducted at academic or other intuitions "for which a federal department or agency has specific responsibility for regulating as a research activity 945 CFR 46.102(e)). Similarly, researchers or institutions applying for funding at the European Union (EU) have to comply with ethical guidelines set out by the EU (European Commission, 2013).

As a result, companies – or any institution – that do not seek such funding or operate outside the institutional boundary of those that do are, therefore, not legally bound to certain ethical regulations (Benbunan-Fich, 2017). There are many forms of experimentation, for example, corporate 'A/B testing' (Benbunan-Fich, 2017), traffic experimentation (Richter et al., 2001; Svensson & Hansson, 2007), the testing of self-driving cars on public roads (Stilgoe, 2020), experiments with predictive policing (Amnesty, 2020) that might benefit from ethical review, yet are not the subject of human research ethics regulation, since the investigators are not tied to regulatory commitments to the same degree as researchers working at a university. A famous example of this was the Facebook Emotional Contagion study, in which researchers at Facebook, in collaboration with Cornell University, studied how emotions spread among users of the platform. The Cornell University researchers had sought IRB approval for this study, but since data collection was technically done independently from Cornell by Facebook researchers before their involvement, the Cornell review board judged that no review was necessary (Flick, 2016).

When conducted by parties tied to federal funding, what matters in terms of regulatory scope is whether a certain activity meets the definition of research or human subjects. If a certain activity falls outside the definitions, research ethics regulation currently does not apply. Research is commonly defined as an activity characterized as a systematic investigation with the intention to develop generaliza-

ble knowledge (US HHS 45.CFR.46). Here, I follow the Harvard Committee on the Use of Human Subjects, (CUHS) which defines investigation as:

> "a methodical procedure and plan, is theoretically grounded, and specifies a focused and well-defined research problem or question, is informed by the empirical findings of others, is analytically robust, and provides a detailed and complete description of data collection methods" (Harvard CUHS).

Drawing again from the Harvard CUHS, generalizable knowledge can be defined as information that:

> "is expected to expand the knowledge base of a scientific discipline or other scholarly field of study and yield … results that are applicable to a larger population beyond the site of data collection or the specific subjects studied [or] results that are intended to be used to develop, test, or support theories, principles, and statements of relationships, or to inform policy beyond the study" (Harvard CUHS).

Human subjects, as defined by the Common Rule, are any living individuals about whom an investigator conducting research:

> (i) "Obtains information or bio-specimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or bio-specimens; or
> (ii) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable bio-specimens" (US HHS 45.CFR.46)

Different studies conducted in or with live-in laboratories can both fall within and outside the scope of these definitions. For example, university researchers studying how nudging technologies influence residents into a more sustainable behavioral pattern would fit all the definitions above. Applying a new type of heat isolation in the walls in the live-in laboratory might be classified as research but not as human subject research. An example that fits no definition is interviews by a local newspaper with residents on how they enjoy their stay.

A problem emerges in which certain types of research and the live-in laboratory itself as a platform for experimentation stay out of shot or regulatory obligations. Letting persons live in homes that are under contentious experimentation is not research in itself. It is the creation and demarcation of every ready and available subject population that can be exposed to a series of overlapping experiments that involves and impacts them to various degrees, which may or may not fall within the defined scope of human subject research. However, we cannot treat their involvement in research that falls within or outside this scope as separate. Take the

aforementioned example of the researchers testing new forms of insulation in the walls of the LIP. Even when subjects are not directly involved in data collection, when this intervention turns out to not work or be toxic, it will be residents who are directly affected.

In recent years, there has been increased scholarly attention to what justifies the boundaries of research ethics regulation (Hansson, 2011; Wilson & Hunter, 2010). For example, the rise of company-sponsored online experimentation has received scholarly attention because these practices are not covered by research ethics regulations yet pose similar ethical concerns for subjects to scientific research (Benbunan-Fich, 2017). Similarly, the residents of LIPs are a vulnerable research population that, due to the intertwinement of their residency with participation and the costs of withdrawing, might not be as well-suited to protect their interests as other research participants might be. If we allow people to participate in live-in laboratories, this participation should be informed by the constraints and influences placed upon residents and the importance of the right to withdraw.

6.5. Research and The Right to Withdraw

Why should we care about the right to withdraw? In this section, I provide two reasons that outline the ethical foundation of a participant being free to withdraw without penalty. First is to make an appeal to codified research norms as the source of an experiment's moral permissibility and hold that an experiment's permissibility is determined by its capacity to comply with research ethics guidelines and, subsequently, be deemed acceptable by ethics commissions or institutional review boards (IRB's). I call this the institutional defense. Afterward, I will provide a moral defense grounded within bioethical principlism.

6.5.1. An Institutional Defense of The Right to Withdraw

The institutional defense holds that an experiment's ethical permissibility is grounded in the judgment or authority of a research ethics committee or institutional review board (IRB). Such a view is, for example, articulated by McNeill, writing that:

> "the principle method for ensuring that human experimentation is ethical is to require researchers to have their proposals for experimentation on human subjects approved by a research ethics committee" (McNeill, 1993, p.1).

Such approval is generally contingent on whether an experiment design complies with internal, national and international documents that set global research practice standards for the permissibility of an experiment's design, process, or effects on human subjects.

The justification for extending the RTW to research participants is hence grounded in their presence in those documents that set the global convention of ethical research, which influences IRB's approval. The RTW is such a right. In fact, Edwards has stated that a reference to the RTW "is now included almost mechanically by researchers and research ethics committees alike" (Edwards, 2005, p. 114).

Let us turn to influential contemporary sources that explicitly mention the RTW. For example, The Declaration of Helsinki (1964, latest revision in 2013) from the World Medical Association states in its 26th principle that:

> The potential subject must be informed of the right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal (2013).

Similar definitions appear in the 'International Ethical Guidelines for Health-related Research Involving Humans' (1993, latest revision in 2016) by the Council for International Organizations of Medical Sciences (CIOMS), which was founded by the WHO and UNESCO:

> "Participants have a right to withdraw at any point in the study without retribution" (2016, 33), and "the individual is free to refuse to participate and will be free to withdraw from the research at any time without penalty or loss of benefits to which he or she would otherwise be entitled (Guideline 9)" (2016, 103).

Similarly, The Belmont Report (1979), which was drafted by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in the aftermath of the Tuskegee Experiment Scandal, mentions that a prospective participant should be presented with:

> "A statement offering the subject the opportunity to ask questions and to withdraw at any time from the research" (National, 1979, p. 6).

While the Belmont Report does not specifically mention the fact that participant has a right to withdraw without reprisal, it does so de facto by denouncing "unjustifiable pressures" which urge a course of action for a subject" (idem, p. 7), examples of which include "threatening to withdraw health services to which an individual would otherwise be entitled" (idem, p. 8).

However, holding institutional research norms as the grounds on which we should judge an experiment's ethical permissibility might not be convincing. This defense merely shows that LIP is not in accordance with current institutional guidelines about ethical experiments. While an interesting conclusion, this argument might be too conventional to provide a satisfactory ground on which to judge the permissibility of an experiment. Indeed, history is filled with examples of experiments or trials on human subjects that did receive ethical approval by IRBs but did later turn out to be highly problematic. Often, a major scandal must occur before any reform in ethics codes is seriously undertaken, and what might be impermissible now would have been permissible several decades ago (McNeill, 1993). To satisfy this concern, in the next section, I will aim to provide a moral defense of the claim that the RTW is a necessary condition for an ethical experiment grounded within bio-ethical principilism (Beauchamp, 2016).

6.5.2. A Moral Defense of the Right to Withdraw

Principlism in bioethics arose in the 1970s through two major works - the Belmont Report and the 'Principles of Biomedical Ethics' by Beauchamp and Childress (Beauchamp, 2016). It aimed to ground the conduct of biomedical research on human subjects not on professional conduct but on moral principles. Principlism offers a practical, pluralistic tool for bio-ethical decision making, sidestepping high moral theory and providing an intuitive framework of, in the words of Beauchamp, "general guidelines that condensed morality to its central elements" (Beauchamp, 1995, p. 181). These principles are 'respect for autonomy,' 'beneficence,' 'non-maleficence,' and 'justice.' (Beauchamp & Childress, 1994). Non-maleficence was created as a separate principle by Beauchamp and Childress and was, in the Belmont Report, understood to be included under the principle of beneficence.

The aforementioned four principles do not constitute all of morality, but according to Beauchamp, a selection is necessary for the construction of a normative framework for biomedical ethics (Beauchamp, 1995). The principles are understood to not take precedence over another and are binding unless they conflict with other moral principles, which allows them to be overwritten by other moral considerations (Beauchamp, 1995). While developed in the context of biomedical ethics, the principles have since been applied to structure ethical decision-making in many other research domains (Bredenoord, 2018).

Beauchamp holds that certain principles are necessary for promoting human flourishing. Beauchamp claims that there is a "tendency for the quality of people's

lives to worsen," which certain principles help to counteract (Beauchamp, 2016, p. 9). What justifies the principles is simply that they are those norms that are effective, or as Beauchamp puts it:

> "Best suited to achieve the objective of morality, which is the promotion of human flourishing by counteracting human circumstances in interactions with others that cause the quality of people's lives to worsen" (Beauchamp, 2016, p. 9).

I argue that the right to withdraw can be understood as a prerequisite to a participant's liberty, understood to be a necessary condition to the principle of autonomy. Beauchamp holds two concepts to be necessary conditions for a person's autonomy; liberty and agency (Beauchamp, 2016). The focus of my argument is on liberty, which Beauchamp defines in the 'negative' sense as "the absence of controlling influences" (Beauchamp, 2016, p. 5). I hold that the function of the RTW is to realize this notion of liberty by providing mechanisms to participants that prevent said controlling influences on their liberty. In other words, the function of the RTW is to prevent investigators from placing constraints on withdrawing from research in order to safeguard a participant's liberty.

However, here we run into two problems. First, we aim to defend not only the ethics of a participant, the right to withdraw from research but also that they have the right to withdraw from research without penalty or loss of benefits. The question stands whether penalizing a research participant can be defined as constraining a participant's liberty. I believe that penalizing does, in fact, constrain a participant's liberty to withdraw from research. Here, I understand constraints on a person's liberty in a 'broad' sense, which includes both intended and 'unintended' restrictions (Carter, 2003). Penalties pose a certain obstacle or interference to people. Exiting costs of a live-in laboratory might not be an intended policy, yet even if unintended, they can constrain a participant's freedom since penalties are a controlling influence. The threat of penalties might deter people from certain actions and urge a certain course of action. I argue in the next chapter that we can understand the costs of withdrawing in an LIP as (potentially unintended) penalties.

A second challenge that we encounter is that all principles in principlism, including respect for autonomy, have 'prima facie' standing (Beauchamp & Childress, 2001). While we can imagine many scenarios in which other principles outweigh a participant's autonomy, in the case of the LIP, there are no good overriding moral reasons that justify the constraint of a participant's liberty that urge them to stay in the LIP experiment. Imposing such costs does not benefit the participant. In fact, it

harms them. Neither is the research of such immediate societal impact or danger that keeping participants in the experiment could be justified based on protecting others from harm. Several authors have argued on this basis that in certain experiments, such as infectious disease studies (Fernandez Lynch, 2020) or xenotransplantation (Spillman & Sade, 2007), we should not award participants the RTW. However, the LIP conducts no research that poses a danger to society when its participants withdraw.

Finally, placing penalties on withdrawal in this form of experimentation can be considered a unjust distribution of the benefits and costs of research participation. While beneficial to the researcher and innovators, residents do not necessarily directly benefit from successful innovations that are tested in LIP, as might be the case with experimental medical trials in which a patient's health is at stake. Hence, there are no overriding reasons to curb an LIP participant's right to withdraw.

6.6. Do the Costs of Withdrawing Qualify as a Penalty or Loss of Benefit?

Earlier, we showed that withdrawing from an LIP can cause financial and mental strains for participants due to the fact that they need to find a new home. This process can be strained due to external factors such as the availability of housing, the capacity of participants to obtain housing, and the uncertain limbo state between withdrawing from the experiment and moving into a new home. In this section, I argue that we can consider such consequences as penalties or losses of benefits that a person is otherwise entitled to.

A common definition frames a penalty as a punishment in reaction to an individual who has violated a rule. In other words, it is a deliberate action in reaction to a violation with the intent to punish. Feinberg argued that while penalties and punishments are both "authoritative deprivations for failures," their difference lies in their level of 'expressiveness', with punishment having a "symbolic significance largely missing from other kinds of penalties" (Feinberg, 1965, p. 400).

However, this intentional notion of a penalty only allows us to qualify the negative consequences that are intentionally given in reaction to said participant withdrawing their research participation consent as penalties. While I do not want to exclude this possibility, I aim to conceptualize penalties without relying on intention since the design and operation of the live-in lab generate a certain environment from which certain negative consequences arise upon withdrawal rather than from the intentions of the investigators.

How about a loss of benefits that a participant is otherwise entitled to? Schaefer and Wertheimer maintain that what a participant is entitled to is limited to those things that were promised to them on either the completion or partial completion of research (Schaefer & Wertheimer, 2010). Benefits are akin to compensation promised. So, not providing a benefit to a participant who was part of the research participation does not necessarily mean that a participant is losing out on something that they would be entitled to, as long as a participant receives what they were promised for the work that they did. If a participant receives less than promised, then they would be penalized. This seems in line with the CIOMS guidelines, which recommend that those who withdraw from research themselves should be compensated proportioned to the part they have completed. In this case, a participant is not entitled to the full amount (CIOMS, 2016).

However, if live-in lab participants were promised a new home upon withdrawal and they would not receive it, this would then constitute a loss of benefits that a participant is otherwise entitled to. Again, this seems to be a possible scenario; however, I do not wish to build my defense of this contingency. So, it seems that this notion is also not helpful in framing the possible costs as penalties.

A potential strategy is to consider the protective intent with which the RTW was introduced into research ethics guidelines. The original inclusion of 'without reprisal' is linked by Melhalm and colleagues to the needs of an important research demographic of (bio)medical research: patients (2014). They argue:

> "Because many participants are recruited by virtue of being patients, in order for their choice to be meaningfully voluntary, there must be an assurance that abstaining or withdrawing will not compromise their current and future clinical care" (Melhalm et al., 2014, p. 3).

The 'without penalty' quality of the RTW – and the 'voluntariness' it was aimed to protect - was therefore originally included to compensate for a patient's natural vulnerability, preventing the threat of losing out on relevant care upon withdrawal would urge a certain course of action of patients. To clarify, the RTW does not ensure that participants can participate and have a right to withdraw unscathed. After all, research often involves certain justifiable risks. However, what the RTW does aim to protect is that withdrawing from participation in itself does not leave participants worse off than they were before participating.

If we conceptualize penalties in the case of the RTW as reductions of a pre-experiment baseline due to withdrawal, then we can categorize the negative consequences of withdrawing from a live-in lab as penalties. Since withdrawing itself, not

the risks that a participant endures during the experiment leaves a participant arguably worse off than before they participated. This definition circumvents the intentional problem and the promise problem by not making the definition of penalty contingent on an intentional character and not focusing on defining a penalty in relation to what a participant was promised for (part of) their research participation. Instead, it focuses on a comparison of a participant's baseline previous to LIP participation and how withdrawing itself penalizes a participant compared to this pre-experiment baseline.

6.7. Do the Costs of Withdrawing Qualify as Unjust Controlling Influences?

Earlier, I argued that the RTW safeguards a participant's liberty and that the potential costs of withdrawing from an LIP can be considered a penalty. This section argues that the potential costs of withdrawing in a LIP could be categorized as a constraint on a participant's liberty since they pose *controlling influences*. This is problematic in regard to the idea that liberty is a necessary condition for the principle of autonomy. If a participant in an experiment lacks that liberty for no apparent justifiable overriding reason, such an experiment should be considered morally suspect.

There exists a strong link between the activity of research participation and the notion of voluntariness. Not only is participation in research understood to be voluntary (Levine, 1996), but also a participant's agreement to participate in research – their informed consent – rests on voluntariness. The RTW can be understood as an essential part of informed consent (Nelson & Merz, 2002). Hence, just as a participant's informed consent is only understood to be meaningful if a participant gave their informant consent voluntarily, so is their right to withdraw. As mentioned earlier, the original inclusion of the 'without penalty' clause was motivated to ensure that a potential participant's choice to participate and stop participating would be meaningfully voluntary given their vulnerable status. In order words, for informed consent and the RTW to be meaningful, one needs to be able to exercise it voluntarily.

This paper outlined a number of financial, psychological, and social costs, which are amplified by certain context-dependent factors that affect a participant withdrawing from the LIP. The prospect of having to endure the aforementioned cost urges a certain course of action for a participant; namely, they influence one's decision-making concerning whether they would withdraw from the experiment.

This scenario seems akin to other situations in which a person is awarded a certain right, but external factors inhibit the right from being freely exercised if the person does not possess a reasonable capacity to overcome those factors. If a person has a right to vote but risks losing their job and hence livelihood when they have to stand in line all day in order to exercise that right, one might be pressured into a certain course of action, namely not to go vote. Similarly, research showed that while US citizens have a federal right to abortion and US states are limited in their capacity to prohibit them, abortions can be discouraged nonetheless through what Johnson and Bond call "a variety of coercive and non-coercive policies that might operate to alter the utilities associated with having or providing abortions" (1980, p. 106).

Imagine a participant who wants to terminate their research participation. They realize that this would mean they have to move out of their house and that this will be financially and emotionally costly for them. Perhaps they do not have the funds to find alternative housing. Such considerations about future potential costs can be reasonably assumed to influence some LIP participants into either postponing their withdrawal or forgetting about the idea altogether. Whether participants necessarily are aware of those costs or consider them to be of no influence is irrelevant to their existence, being a possible influence on those participants who do consider and are influenced by them. So, the cost functions as a pressure that urges a certain course of action, which is to not withdraw. As argued earlier, we have no reason to assume that the costs of withdrawing from the LIP qualify as potential justified pressures. In other words, a participant of the LIP is unable to freely exercise their right to withdraw.

6.8. Conclusion

In this paper, I have identified the negative consequences of withdrawing from a laboratory that is also a participant's home and argued that these consequences are morally problematic when held against an appropriate normative research framework. Specifically, participants are unable to withdraw from research without the (threat of) losing their homes. This strains research participants' ability to exercise the right to withdraw, which they are awarded based on the virtue of them being research participants. I have grounded the ethical justification of the RTW in both institutional convention and biomedical principlism as a mechanism for realizing a participant's liberty, understood as a necessary condition for the value of autonomy.

I have shown that the negative consequences of withdrawing from an LIP can both be categorized as a penalty and a controlling influence, meaning LIP participants are not able to exercise their RTW freely and without penalty.

However, the point of this paper is not to claim that live-in laboratories are an unacceptable research methodology. Instead, the aim is to highlight that an intimate intertwining of a research participant's daily and experimental life facilitates a problematic violation of established ethical norms. Experiments within living society raise the question of whether participants are able (and should be able to) withdraw. Yet, how can a participant withdraw from real life? This paper underpins the necessity for investigations into the normative boundaries of urban experimentation that affect human beings. In this last section, I want to briefly propose such a boundary: restrict live-in laboratory use to temporary residents.

Let us first explore the alternative solution: cover potential costs that withdrawing imposes on participants through compensation. For example, participants could be promised that if they withdraw, similar and adequate housing will be provided for them and that they will be assisted financially in the moving process. If we assume that all costs of withdrawing are nullified through investigators' efforts, research participants would arguably not be penalized and influenced in their decision to withdraw from the LIP. In fact, the student studios at the Green Village, where the LIP Dreamhus are also located, already have a relocation policy in place.

However, this strategy does have its downsides. Namely, it commits the investigators to the use of the LIP but leaves other potential problematic aspects of the experimental apparatus unresolved. For example, it remains unclear why it would be epistemically beneficial to have participants live for such a long duration in a laboratory setting. LIPs might also be problematic independent of their use since, by virtue of their design, they do not allow participants to realize their privacy.

A second solution does not face such problems. This strategy proposes to untangle the interwoven relations between a participant living their daily life and being part of an experiment. By prohibiting investigators or participants from making a live-in laboratory a Lived-in Place – a permanent residency – and instead limiting their presence to temporary visits, like a Visited Place, many of the above-mentioned problems can be prevented. Participants would not need to worry about any negative consequences of withdrawing from the LIP since they could simply leave their human vivarium and go home.

7. Conclusion

This thesis provided a comprehensive analysis of research with technology under real-world conditions, its ethically relevant aspects, and the shortcomings of applying existing research ethics principles and norms to these research types, which I further illustrated in two specific cases: real-world AI research and live-in laboratories. Taking these existing research ethics principles and norms as a serious starting point for analysis, the thesis showed that there are obstacles to their application to real-world research for reasons shared with all relevant cases of real-world research. From identifying these gaps, I argued, we can identify pointers for developing new research ethics guidelines for real-world research.

To reiterate, the main research question that this thesis aimed to answer was:

> RQ: *In which ways do common ethical challenges emerge in research under real-world conditions, and in what ways do research ethics principles and norms fail to account for these ethical challenges?*

I approached this question through an applied philosophical approach. I drew from philosophical insights, particularly from research ethics, and a wide variety of cases of real-world technology research in order to analyze their common characteristics and challenges. In doing so, I laid the foundations for the ethics and governance of real-world research and stressed a need to develop new research ethics that address the shared characteristics of real-world research and identify challenges that such an approach should consider. The thesis proceeded as follows:

First, in Chapter 2, I laid the groundwork for a comprehensive ethical analysis of real-world research. Specifically, I addressed the question:

> RQ2: *Are there unifying and ethically significant features common to all relevant examples of real-world research?*

I answered this question in the affirmative. In doing so, it provided a justification for the focus and scope of this thesis. Despite the apparent diversity of real-world research, which lends itself to vivid contrast, I provided a novel account of what unifies seemingly diverse forms of real-world research, ranging from self-driving car tests to online A/B tests, and can be considered a morally salient feature.

That feature is 'coupling.' Coupling is defined as when two potentially independent options are 'coupled,' meaning you cannot choose (or reject) one without the other. This might occur, for example, when research is conducted under real-world conditions in a public street, and a person can no longer engage with this environment – walk that street — without being subjected to the research. Before the research was set up, people had the option of 'going to the street' without the option of 'being a subject.' After the experiment is set up, both options are coupled, and one can no longer accept or reject one without the other. I argued that this coupling can give a reason for moral concern, depending on the moral salience of (and the interplay between) the options themselves that are coupled, as well as the degree of control one has over the coupling in the first place.

In Chapter 3, I discussed the general absence of ethical governance in real-world research, especially compared to scientific research. Specifically, I asked:

> RQ3: *What, if anything, justifies the lack of research ethics governance for real-world research in comparison to scientific research?*

I examined and rejected four possible justificatory reasons that might justify this status quo, and I argued that asymmetrical research ethics demands create the opportunity for the potential avoidance of research ethics burdens by placing research activities outside the current scope of research ethics regulation, which comes at the expense of those whom these demands aim to help protect. Thus, we have no good reasons not to develop a research ethical governance for real-world research.

In Chapter 4, I analyzed the application of paradigmatic research ethics norms to real-world research. Specifically, I aimed to answer the following question:

> RQ4: *What challenges emerge in applying paradigmatic research ethics norms to real-world research?*

Specifically, I focused on what I call the identification problem of real-world research. I argued that real-world research often involves significant uncertainty and difficulty in identifying its exact scope or reach. This is problematic since many paradigmatic norms assume an investigator can identify those persons to whom ethical obligations are owed (this, of course, does not necessarily mean that these demands no longer apply). As a consequence, these norms are impractical to comply

with. Examples include providing information about the research, informed consent, and a just distribution of research participants.

To clarify, in Chapter 4, I do not argue that these norms *should* be the basis for ethically evaluating real-world research. Rather, it shows that were we to hold real-world research to these norms – as some scholars have done in similar cases – then these norms would be difficult or impossible to uphold (and thus render much of real-world research unethical if it were argued that these norms should indeed apply, given an absence of overriding reasons). So, while Chapter 3 argued that real-world research practices should be subject to a consistent research ethics governance, in Chapter 4, I show that if we base the ethical content of this governance on paradigmatic research ethics norms, this would require severe research redesigns to ensure such norms can be upheld, which might impact their epistemic value.

In Chapters 5 and 6, I analyzed two distinct case studies of real-world research: real-world AI research (Chapter 5) and live-in laboratories (Chapter 6). In doing so, I aimed to accomplish two goals. First, I will show how the issues raised in this dissertation play out on a case basis. Second, to show that even though, at face value, these examples have stark contrasts, they both present similar ethical concerns due to them being real-world research.

First, in Chapter 5, I provide an analysis of the ethical challenges of real-world research with LLMs and generative AI. Specifically, I asked:

> RQ5: *What challenges arise when we evaluate real-world AI research with paradigmatic research ethics principles?*

I argued that despite its potential epistemic value, real-world AI research faces challenges in meeting research ethics principles influential to research ethics standards — non-maleficence, beneficence, respect for autonomy, and distributive justice — and that these challenges are exacerbated by absent or imperfect current ethical governance.

In Chapter 6, I continued my focus on the ethics of real-world research on a case-study basis by turning to live-in laboratories: homes built as experimental living environments to test the performance of novel technologies on their residents. Live-in laboratories embody the major themes of this thesis. First, real-world research 'couples' itself with daily life. Live-in laboratories, 'couple' in a clear sense of daily life with research participation. Secondly, paradigmatic research ethics norms are difficult or impossible to uphold in real-world research. I showed how these two

themes come together in live-in laboratories by analyzing how the 'right to withdraw' — a research ethics norm that grants research subjects the ability to withdraw from research without penalty or coercive influences in order to safeguard the voluntary status of research participation – conflicts with this real-world research practice. Specifically, I asked:

RQ6: *How, if at all, does the right to withdraw conflict with live-in laboratory research?*

The chapter argued that live-in laboratory research conflicts with this paradigmatic norm and concludes that if we were to take the right to withdraw seriously, then the practice of coupling a participant's main residence to research participation would be ethically problematic.

7.1. Key Findings

The aim of this thesis is to make an impact both in and beyond the fields of research ethics and technology ethics and to also be of interest to a wide variety of scholars, practitioners, and government facilitators interested in (the research ethics and governance of) research under real-world conditions. In doing so, this analysis aimed to bridge the theoretical and the practical in order to lay the groundwork for the ethics of real-world research and identify challenges and areas of attention that such a framework should take into account. Four considerations emerge.

First, one key finding of this dissertation is its novel account of coupling as a common moral characteristic of all real-world research. Real-world research 'couples' research participation in daily life in a way that many other forms of research do not. This coupling is salient since it presents a cost to avoiding research participation (or becoming a research bystander). My account of coupling suggests that we should view real-world research that couples weighty, significant choices, like where one lives with an experiment, with more suspicion than research that couples more insignificant choices, such as which street one visits in a city. Additionally, the more negative moral salience the experimental option presents, the more we should view the resulting real-world research with suspicion.

Second, as real-world research couples itself intimately with our daily lives, there is a pressing need for ethical governance. However, real-world research lacks ethical governance, and attention to this issue has been fragmented. This absence is important since research ethic governance, at least theoretically, aims to provide

guardrails and ensure that research is conducted in accordance with ethical principles or norms. Thus, the absence of research ethics governance for real-world research comes at the potential cost of those the research potentially affects. My analysis shows that the lack of research ethics governance is difficult to justify and that we, thus, have good reasons to develop a research ethics governance for real-world research.

Third, despite the need for ethical governance, this does not mean we must draw from the same ethical content as found in existing research ethics guidelines. This is because a close analysis of the characteristics of real-world research reveals that a direct application of paradigmatic research ethics protocols and norms is problematic. One reason for this is that since real-world research is conducted in largely uncontrolled environments, real-world research has a problem with identifying who it involves and affects. So, if we maintain that researchers need to comply with paradigmatic research ethics norms, the identification problem of real-world research suggests that many such norms that focus on the individual will be difficult to uphold in real-world research environments.

This has far-reaching implications for our efforts to conduct and evaluate research under real-world conditions since it would mean that we cannot neatly apply paradigmatic research ethics norms to real-world research. As I pointed out, this thesis does not argue that paradigmatic research ethics norms *should* be the basis for ethically evaluating real-world research. Rather, it shows that if we were to hold real-world research to these norms – as some scholars have done in similar cases – (and we cannot find any overriding reasons), they would render much of real-world research ethically problematic.

Such a conclusion could yield two responses. Either we would have to severely amend real-world research practice to ensure they can uphold such paradigmatic research ethics norms. This might result in undermining the potential practical and epistemic benefit of real-world research. Alternatively, we would have to find overriding reasons for all the norms that they are in violation (of which I showed the range in Chapter 4). After all, within the research ethics literature, principles and norms are often not taken to be absolute and can be overridden. However, even if we accept particular overriding conditions, as I showed in Chapter 3, this will leave out research cases that do not meet these conditions. Additionally, overriding reasons should be found for as many as the paradigmatic research norms affected. This could prove troublesome since it involves a potentially long list of affected norms. A better and third solution is not to wholly use existing ethics guidelines and

instead develop research ethics governance for real-world research that accounts for the specific ethical challenges of real-world research and contains ethical content that aligns with the realities of real-world research practice.

Fourth, real-world research can present problems regarding the distribution of research ethics responsibilities. Real-world research often involves transdisciplinary collaborations between academic, public, and private entities, for example, organized in so-called 'real-world laboratories,' 'test beds,' or 'living labs.' Since many non-scientific parties are not held to the same research ethical standards, this creates problematic differences between the research parties involved. Whatever research ethics norms for the real world we settle on, these need not only to be appropriate to real-world research but, when conducted in transdisciplinary research collaboration, also be ideally harmonized across all parties involved in order to avoid regulatory evasion and the diffusion of the responsibilities.

To summarize, this thesis does not claim that real-world research is an unethical or unacceptable research methodology. Instead, it has shown through a comprehensive analysis of real-world research that real-world research is unified in an intimate and morally salient intertwining of people's daily and experimental lives and that this facilitates a problem for paradigmatic research ethics norms and principles. It prompts further investigations into the normative boundaries of real-world research and the development of new ethical codes. In the absence of ethical guardrails for real-world research, parties engaged in real-world research should not be left alone in this endeavor. Ethicists should continue to draw attention to the ethical challenges of real-world research and aid researchers in the public and private sectors to navigate those moral challenges.

7.2. Avenues for Future Research and Limitations

This dissertation covers a lot of ground. Nevertheless, several gaps persist. Here, I will list three prominent avenues for future research. After, I outline several limitations of this dissertation.

7.2.1. Avenues for Future Research

A prominent avenue is developing a new code of ethics for real-world research. This dissertation did not aim to provide an all-encompassing action-guiding ethical framework for ethical real-world research. However, it repeatedly stressed the need for one and gave several recommendations for its development. A critical reader

might wonder why I have not developed such a framework so in this thesis. However, as I outlined in the introduction, I have argued that before we can develop such a framework, particular groundwork is necessary, which this thesis addresses. Particularly, a clear case needed to be made that (1) we can consider real-world research a unified and morally interesting unit of analysis for which to develop research ethics guidelines specifically, (2) in which ways common ethical challenges emerge in research under real-world conditions, and (3) what ways do existing research ethics frameworks fail to account for these ethical challenges and thus need to be accounted for in a new research ethics framework.

I have addressed these points by offering a comprehensive analysis of the characteristics of real-world research, which is a unified and morally interesting unit of analysis due to, at least, my account of coupling and which is in need of research ethics governance. My analysis further revealed that paradigmatic research ethics norms are problematic to uphold in real-world research due to what I have called the identification problem, and, thus, that in considering new research ethics of ethics for real-world research, we need to be sensitive to its specific challenges and characteristics. Therein lies an obvious avenue for future research: to develop such a framework.

A second avenue for future research concerns a further analysis of coupling, specifically in relation to the notion of personal control and philosophical theories of freedom and power. To briefly recap what I mentioned in the discussion section of Chapter 2, coupling seems particularly morally salient when one has no control over the coupling. I outlined that a promising lens through which to approach this issue is through neo-republican accounts of freedom. Unlike liberal traditions that define freedom as freedom from *actual* interference (Berlin, 1969), neo-republicanism defines freedom as being free from (or the absence of) the *potential* interference of arbitrary (or *dominating*) exercises of power (Lovett, 2010). Recently, scholars have used this theory to analyze, for example, digital nudges (Capasso, 2022)— features of user-interface design, like dark patterns, that guide people's behavior in online choice environments (Weinmann et al., 2016)— or certain forms of risk impositions (Maheshwari & Nyholm, 2022). These dominating digital nudges and risk impositions are wrong when and because they actualize an influence that is not *controlled* by the people they target (Capasso, 2022). Such accounts offer a promising starting point for future research into how coupling intersects with personal control and freedom and when it might dominate.

A third avenue for further analysis concerns the following question: while real-world research might be unified by coupling, does all real-world research couple in the same way? As I mentioned in the discussion of Chapter 2, it seems that different real-world research environments might have seemingly different opportunities to decouple options. To reiterate, decoupling refers to separating coupled options into two or more distinct options that can be chosen independently. Specific avenues for personal control are more easy to offer, often in online real-world environments than in 'physical' real-world environments. A physical location like a public square can't appear to the same person in two different ways in the way that a website can present different alterations to various persons, each simultaneously interacting with a distinct version of the same website. Specific individual interactions (for example, interacting with a specific biometric border gate in an airport[43]) could sometimes be 'decoupled' from other interactions in the same physical space. This insight provides an interesting avenue for a more detailed taxonomy of coupling and real-world research.

7.2.2. Limitations

Turning to the limitations, the first limitation builds on the fact that this dissertation does not answer the question of whether we have reason to hold non-scientific entities, such as governments, corporations, engineers, hospitals, NGOs, etc., to the *same* research ethical demands as their counterparts in science as a matter of morality. This is important since, if this were the case, it would offer clarity regarding the moral obligations we should hold these researchers to. However, given the argument in Chapter 4, such a conclusion would also provide us with moral reasons to severely amend real-world research practices to ensure they can uphold such paradigmatic research ethics norms. However, this, in turn, might impact the potential practical and epistemic benefits of real-world research, bringing about particular costs in this regard. Alternatively, we would have to find overriding reasons for all the norms that they are in violation of. However, I showed the range and challenges of this approach in Chapter 4.

---

[43] In 2017, Schiphol Airport in Amsterdam conducted a test with biometric boarding gates, aiming to make boarding more efficient. Rather than conducting the test airport-wide, the test was conducted at a specific gate within the airport. Participation was voluntary and only participants would use the experimental facial recognition gate (Schiphol, 2017).

A second limitation concerns my argument in Chapter 3. In this Chapter, I rejected four potential reasons that might justify the lack of research ethics regulations for real-world research and argued that the current situation of unequal demands can be exploited by circumventing ethical demands at the cost of a person's protection. However, if my argument is inherently a comparative one, this places my claim (that we should extend research ethical governance to real-world research) on a relatively weak foundation. This is because, as I mentioned in the discussion section of Chapter 3, the upshot of these claims can swing both ways. Either this argument can justify reducing the regulatory burden on currently regulated research to bring it closer to unregulated research practice (much real-world research), or it can be used to increase the regulatory burden on currently unregulated research (much real-world research). I mentioned that I believe a stronger case can be made to bring currently unregulated research more in line with currently regulated research and pointed to some literature (see, for example, Hansson (2011)) which had done work on this. However, I have not defended this; hence, it remains an important limitation that needs further attention to resolve. Additionally, I consider a limited range of reasons that might justify inconsistencies between the presence of research ethics governance for real-world research and other forms of research. Hence, more work is needed to consider a more fuller set of possible justifying reasons.

References

Ada Lovelace Institute. (2022). Looking before we leap: Ethical review processes for AI and data science research. Available at: https://www.adalovelaceinstitute.org/report/looking-before-we-leap/

Alavi, H. S., Lalanne, D., & Rogers, Y. (2020). The five strands of living lab: A literature study of the evolution of living lab concepts in HCI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *27*(2), 1-26.

Amnesty International. (2020). *Netherlands: We sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands.* Retrieved March 2022, from https://www.amnesty.org/en/documents/eur35/2971/2020/en

Ansell, C. (2012). What is a "democratic experiment"? *Contemporary Pragmatism, 9*(2), 159–180.

Ansell, C. (2019). "Coping with Conceptual Pluralism: Reflections on Concept Formation." *Public Performance & Management Review* 44 (October):1–22. https://doi.org/10.1080/15309576.2019.1677254.

Ansell, C. K., & Bartenberger, M. (2016). Varieties of experimentalism. *Ecological Economics*, *130*, 64-73.

Ansell, C., & Bartenberger, M. (2017). The diversity of experimentation in the experimenting society. In *New Perspectives on Technology in Society* (pp. 36-58). Routledge.

Arntzen, S., Wilcox, Z., Lee, N., Hadfield, C., & Rae, J. (2019). Testing innovation in the real world. *London: Nesta.*

Baccarne, B., Schuurman, D., Mechant, P., & De Marez, L. (2014). The role of urban living labs in a smart city. In *XXV ISPIM Innovation Conference*.

Barmpounakis, E., & Geroliminis, N. (2020). *On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment.* Transportation research part C: emerging technologies, 111, 50–71.

Benbunan-Fich, R. (2017). The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation. *Research Ethics*, *13*(3-4), 200-218.

Benham, B. (2008). Moral accountability and debriefing. *Kennedy Institute of Ethics Journal*, *18*(3), 253–273.

Bean, S. (2010). Beyond research exceptionalism: a call for process redesign. *The American Journal of Bioethics*, *10*(8), 58-60.

Beauchamp, T. L. (1995). Principlism and its alleged competitors. *Kennedy Institute of Ethics Journal*, *5*(3), 181-198.

Beauchamp, T. L. (2016). Principlism in bioethics. *Bioethical decision making and argumentation*, 1-16.

Beauchamp, T. L., & Childress, J. F. (1994). *Principles of biomedical ethics*. Edicoes Loyola.

Beauchamp, T. L. & Childress, J. F. (2001). *Principles of Biomedical Ethics*. Oxford University Press.

Beerbohm, E., Davis, R., & Kern, A. (2020). The democratic limits of political experiments. *Politics, Philosophy & Economics*, *19*(4), 321-342.

Beckers, R., Kwade, Z., & Zanca, F. (2021). The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. *Physica Medica*, *83*, 1-8. https://doi.org/10.1016/j.ejmp.2021.02.011

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Benatar, S. R. (2004). Towards progress in resolving dilemmas in international research ethics. *Journal of Law, Medicine & Ethics*, *32*(4), 574-582.

Benatar, S. R., & Singer, P. A. (2000). A new look at international research ethics. *Bmj*, *321*(7264), 824-826.

Benatar, S. R., & Singer, P. A. (2010). Responsibilities in international research: a new look revisited. *Journal of medical ethics*, *36*(4), 194-197.

Berlin, Isaiah, 1969. 'Two Concepts of Liberty,' in his *Four Essays on Liberty*, Oxford: Oxford University Press: 118–72.

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, 100177. https://doi.org/10.1016/j.caeai.2023.100177

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, *15*(3), ep430. https://doi.org/10.30935/cedtech/13176

Bogers, M., Chesbrough, H., & Moedas, C. (2018). Open innovation: Research, practices, and policies. *California Management Review*, *60*(2), 5–16.

Bommasani, R., Liang, P., & Lee, T. (2023). Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.15007

Boring, E. G. (1954). The nature and history of experimental control. *The American Journal of Psychology*, *67*(4), 573-589.

Bracken-Roche, D., Bell, E., Macdonald, M. E., & Racine, E. (2017). The concept of 'vulnerability' in research ethics: an in-depth analysis of policies and guidelines. *Health research policy and systems*, *15*(1), 1–18.

Bredenoord, A. L. (2018). The principles of biomedical ethics revisited. In *Islamic Perspectives on the Principles of Biomedical Ethics: Muslim Religious Scholars and Biomedical Scientists in Face-to-Face Dialogue with Western Bioethicists* (pp. 133-151).

Brey, P. (2017). Ethics of emerging technology. *The ethics of technology: Methods and approaches*, 175–191.

Brey, P., & Dainow, B. (2023). Ethics by design for artificial intelligence. *AI and Ethics*, 1–13.

Buocz, T., Pfotenhauer, S., & Eisenberger, I. (2023). Regulatory sandboxes in the AI Act: Reconciling innovation and safety? *Law, Innovation and Technology*, *15*(2), 357-389.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Capasso, M. (2022). Manipulation as digital invasion: A neo-republican approach. In *The Philosophy of Online Manipulation* (pp. 180–198). Routledge.

Calo, R. (2013). Consumer subject review boards: A thought experiment. *Stan. L. Rev. Online*, *66*, 97.

Carter, I. (2021). Positive and Negative Liberty. *Stanford Encyclopedia of Philosophy*, Accessed 22 December 2021. http://plato.stanford.edu/archives/spr2012/entries/liberty-positive-negative/.

Chesterman, S. (2020). Artificial intelligence and the problem of autonomy. *Notre Dame J. on Emerging Tech.*, *1*, 210.

Council for International Organizations of Medical Sciences (CIOMS). (2016). *International Ethical Guidelines for Health-related Research Involving Humans*. Accessed 07 January 2022. https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf

Collingridge, D. (1982). The social control of technology.

Colonna, L. (2023). The AI Act's Research Exemption: A Mechanism for Regulatory Arbitrage? In *YSEC Yearbook of Socio-Economic Constitutions 2023: Law and the Governance of Artificial Intelligence* (pp. 51–93). Cham: Springer Nature Switzerland.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., ... & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine*, *3*(1), 141. https://doi.org/10.1038/s43856-023-00370-1

Dainow, B., & Brey, P. (2021). Ethics by design and ethics of use approaches for artificial intelligence. *European Commission DG Research & Innovation RTD*.

David, M., & Gross, M. (2019). Futurizing politics and the sustainability of real-world experiments: what role for innovation and innovation in the German energy transition? *Sustainability Science*, *14*, 991-1000.

de Graeff, N., Pirson, I., van der Graaf, R., Bredenoord, A. L., & Jongsma, K. R. (2023). The identification problem: Defining and delineating the community in field trials with gene drive organisms. *Bioethics*.

DeArman, A. (2019). "The Wild, Wild West: A Case Study of Self-Driving Vehicle Testing in Arizona." *Ariz. L. Rev.* 61:983.

Desposato, S. (2018). Subjects and scholars' views on the ethics of political science field experiments. *Perspectives on Politics*, *16*(3), 739-750

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, *2*(8), 423-425. https://doi.org/10.1038/s42256-020-0219-9

Dobbe, R., & Whittaker, M. (2019). AI and Climate Change: How they're connected, and what we can do about it. *AI Now Institute, Medium*, *17*.

"DreamHûs." n.d. The Green Village. https://www.thegreenvillage.org/project/dreamhus/.

Dusseldorp, M., Does, E., Hillerbrand, R., & Parodi, O. (2024). How to develop a Code of Ethics for real-world labs. *GAIA-Ecological Perspectives for Science and Society*, *33*(4), 397-406.

Edwards, S. J. (2005). Research participation and the right to withdraw. *Bioethics*, *19*(2), 112–130.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*. https://doi.org/10.48550/arXiv.2303.10130

Emanuel, E., Abdoler, E., & Stunkel, L. (2016). Research ethics: How to treat people who participate in research. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 513–523). American Psychological Association. https://doi.org/10.1037/14805-031

Engels, F., Wentland, A., & Pfotenhauer, S. M. (2019). Testing future societies? Developing a framework for test beds and living labs as instruments of innovation governance. *Research Policy*, *48*(9), 103826.

European Commission. Directorate General for Research. (2013). Ethics for researchers: facilitating research excellence in FP7. Publications Office. https://doi.org/10.2777/7491

Europol (2023). *ChatGPT - The impact of Large Language Models on Law Enforcement*, a Tech Watch Flash Report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg

Europol (2024). *AI and Policing – The Benefits and Challenges of Artificial Intelligence for Law Enforcement*. Europol Innovation Lab observatory report, Publications Office of the European Union, Luxembourg

Evans, J., & Karvonen, A. (2010). Living laboratories for sustainability: exploring the politics and epistemology of urban transition. In *Cities and low carbon transitions* (pp. 142-157). Routledge.

Evans, J., Bulkeley, H., Voytenko, Y., McCormick, K., & Curtis, S. (2018). Circulating experiments: Urban living labs and the politics of sustainability. In *The Routledge handbook on spaces of urban politics* (pp. 416-425). Routledge.

Fecher, B., Hebing, M., Laufer, M., Pohle, J., & Sofsky, F. (2023). Friend or Foe? Exploring the Implications of Large Language Models on the Science System. *arXiv preprint arXiv:2306.09928*. https://doi.org/10.48550/arXiv.2306.09928

Fehlmann, T. (2019). Testing artificial intelligence. In *European conference on software process improvement* (pp. 709-721). Cham: Springer International Publishing.

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media+ Society*, *4*(1), 2056305118763366.

Feinberg, J. (1965). The Expressive Function of Punishment. *The Monist* 49(3): 397–423. DOI: 10.5840/monist196549326

Feinberg, J. (1984). *Harm to others* (Vol. 1). Oxford University Press, USA.

Fejerskov, A. (2022). *The global lab: Inequality, technology, and the experimental movement*. Oxford University Press.

Felt, U., B. Wynne, M. Callon, M. E. Gonçalves, S. Jasanoff, et al. (2007). *Taking European Knowledge Society Seriously* (Directorate-General for Research, Science, Economy and Society, Brussels)

Fernandez Lynch, H. (2020). The right to withdraw from controlled human infection studies: justifications and avoidance. *Bioethics*, *34*(8), 833-848.

Flick, C. (2016). Informed consent and the Facebook emotional manipulation study. *Research Ethics*, *12*(1), 14–28.

Flight, S. (2019). Evaluatie Pilot Bodycams Politie Eenheid Amsterdam 2017-2018. *Politie en Wetenschap*.

Fried, B. J., Lagunes, P., & Venkataramani, A. (2010). Corruption and inequality at the crossroad: A multimethod study of bribery and discrimination in Latin America. *Latin American Research Review*, *45*(1), 76-97.

Future of Life Institute (2023). *Pause giant AI experiments: An open letter*. Retrieved from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Future of Life Institute (2023b). *Policymaking in the pause* . Future of Life Institute. Retrieved from https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf

Galič, M. (2019). Surveillance, privacy and public space in the Stratumseind Living Lab: The smart city debate, beyond data. *Ars Aequi, special issue July/August*.

Gelinas, L., Wertheimer, A., & Miller, F. G. (2016). When and why is research without consent permissible?. *Hastings Center Report*, *46*(2), 35-43.

Giziński, S., Kaczyńska, P., Ruczyński, H., Wiśnios, E., Pieliński, B., Biecek, P., & Sienkiewicz, J. (2024). Big Tech influence over AI research revisited: Memetic analysis of attribution of ideas to affiliation. *Journal of Informetrics*, 18(4), 101572.

González-Duarte, A., Zambrano-González, E., Medina-Franco, H., Alberú-Gómez, J., Durand-Carbajal, M., Hinojosa, C. A., ... & Kaufer-Horwitz, M. (2019). II. The research ethics involving vulnerable groups. *Revista de investigación clínica*, *71*(4), 217-225.

GovTech. (2023). *NYC schools working with experts to launch AI Policy Lab*. GovTech. https://www.govtech.com/education/k-12/nyc-schools-working-with-experts-to-launch-ai-policy-lab

Green Village. (2021). Dreamhûs. Accessed 19 December 2021. https://thegreenvillage.org/project/dreamhus/

Green Village. (2023). Nieuwe Buren. Accessed on 4th April 2023. https://www.thegreenvillage.org/nieuwe-buren-dreamhus-2023/

Grimmelmann, J. (2015). The law and ethics of experiments on social media users. *Colo. Tech. LJ*, *13*, 219.

Gross, M. (2018). Real-world experiments as generators of sociotechnical change. In *Energy as a Sociotechnical Problem* (pp. 125–138). Routledge.

Gross, M., & Krohn, W. (2004). Science in a real-world context: constructing knowledge through recursive learning. *Philosophy Today*, *48*, 38.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Hagendorff, T. (2022). A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*, *35*(3), 55.

Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*. https://doi.org/10.48550/arXiv.2303.13988

Hall, T. A., & Hasan, S. (2020). The politics of experimentation. *Available at SSRN 3571296*.

Hansson, S. O. (2004). Weighing risks and benefits. *Topoi*, *23*(2), 145-152.

Hansson, S. O. (2006). Informed consent out of context. *Journal of Business Ethics*, *63*, 149-154.

Hansson, S. O. (2010). Reversing "research exceptionalism". *The American Journal of Bioethics*, *10*(8), 66-67.

Hansson, S. O. (2011). Do we need a special ethics for research? *Science and engineering ethics*, *17*, 21-29.

Hansson, S. O. (2015). Experiments before science. What science learned from technological experiments. *The role of technology in science: philosophical perspectives*, 81-110.

Hansson, S. O. (2016). Experiments: Why and how? *Science and Engineering Ethics*, *22*, 613-632.

Hansson, S. O. (2019). Farmers' experiments and scientific methodology. *European Journal for Philosophy of Science*, *9*(3), 32.

Harbers M, Overdiek A. (2022). Towards a living lab for responsible applied ai. In: Proceedings of the DRS 2022. Retrieved from https://doi.org/10.21606/drs.2022.422

Harrison, G. W. (2005). Field experiments and control. In *Field experiments in economics* (Vol. 10, pp. 17-50). Emerald Group Publishing Limited.

Harvard Committee on the Use of Human Subjects. "How do the federal regulations define research? How Do the Federal Regulations Define Research?" (n.d.). Retrieved May 1, 2023, from https://cuhs.harvard.edu/definition-research

Heiding, F., Schneier, B., Vishwanath, A., & Bernstein, J. (2023). Devising and Detecting Phishing: large language models vs. Smaller Human Models. *arXiv preprint arXiv:2308.12287*. https://doi.org/10.48550/arXiv.2308.12287

Hennaoui, L., & Nurzhan, M. (2023). Dealing with a Nuclear Past: Revisiting the Cases of Algeria and Kazakhstan through a Decolonial Lens. *The International Spectator*, *58*(4), 91-109.

Hodel, D., & West, J. (2023). Response: Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2308.16118*. https://doi.org/10.48550/arXiv.2308.16118

Holm, S. (2011). Withdrawing from research: a rethink in the context of research biobanks. *Health Care Analysis*, *19*, 269–281.

Hudson, J. M., & Bruckman, A. (2004). "Go away": Participant objections to being studied and the ethics of chatroom research. *The Information Society*, *20*(2), 127–139.

Huitema, D., Jordan, A., Munaretto, S., & Hildén, M. (2018). Policy experimentation: core concepts, political dynamics, governance and impacts. *Policy Sciences*, *51*, 143-159.

Humphreys, M. (2015). Reflections on the ethics of social experimentation. *Journal of Globalization and Development*, *6*(1), 87–112.

International Atomic Energy Agency. (2005). *Radiological Conditions at the Former French Nuclear Test Sites in Algeria*. International Atomic Energy Agency.

Intille, S. S., Larson, K., Beaudin, J., Tapia, E. M., Kaushik, P., Nawyn, J., & McLeish, T. J. (2005). The PlaceLab: A live-in laboratory for pervasive computing research (video). *Proceedings of PERVASIVE 2005 Video Program*.

Janssen, H. L. (2020). An approach for a fundamental rights impact assessment to automated decision-making. *International Data Privacy Law*, *10*(1), 76-106. https://doi.org/10.1093/idpl/ipz028

Johnson, C. A., & Bond, J. R. (1980). Coercive and Noncoercive Abortion Deterrence Policies: a Comparative State Analysis. *Law & Policy*, *2*(1), 106-128.

Jouhki, J., Lauk, E., Penttinen, M., Sormanen, N., & Uskali, T. (2016). Facebook's emotional contagion experiment as a challenge to research ethics. Media and Communication, 4(4).

Jurowetzki, R., Hain, D., Mateos-Garcia, J., & Stathoulopoulos, K. (2021). The Privatization of AI Research (-ers): Causes and Potential Consequences--From university-industry interaction to public research brain-drain?. *arXiv preprint arXiv:2102.01648*.

Karvonen, A., & Van Heur, B. (2014). Urban laboratories: Experiments in reworking cities. *International Journal of Urban and Regional Research*, *38*(2), 379-392.

Keeling, G. (2022). Automated Vehicles and the Ethics of Classification. *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*, 41.

Kerrison, S. (2012). Do participants who are harmed in research face too many barriers to claiming compensation? *Research Ethics*, *8*(1), 49–54.

Khanal, S., Zhang, H., & Taeihagh, A. (2024). Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. *Policy and Society*, puae012.

Kimmelman, J. (2020). Why IRBs should protect bystanders in human research. *Bioethics*, *34*(9), 933-936.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. *arXiv preprint arXiv:2301.10226*. https://doi.org/10.48550/arXiv.2301.10226

Kitchin, R. (2016). The ethics of smart cities and urban science. *Philosophical transactions of the royal society A: Mathematical, physical and engineering sciences*, *374*(2083), 20160115.

Klenk, M. (2024). Ethics of generative AI and manipulation: a design-oriented research agenda. *Ethics and Information Technology*, *26*(1), 9.

Kolstoe, S. E., & Pugh, J. (2024). The trinity of good research: Distinguishing between research integrity, ethics, and governance. *Accountability in research*, *31*(8), 1222-1241.

Kramer, A. D. (2012). The spread of emotion via Facebook. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 767-770).

Kramer, A., Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 8788–8790. https:// doi. org/ 10. 1073/ pnas. 13200 40111

Kroes, P. (2016). Experiments on socio-technical systems: The problem of control. *Science and engineering ethics*, *22*(3), 633–645.

Krohn, W., & Weyer, J. (1994). Society as a laboratory: The social risks of experimental research. *Science and public policy*, *21*(3), 173-183. https://doi.org/10.1093/spp/21.3.173

KTH (2020). "Application to live at KTH Live-in Lab" Retrieved March 3, 2023, from https://www.liveinlab.kth.se/en/nyheter/aktuellt/application-to-live-at-kth-live-in-lab-1.971802

Kudina, O., & Verbeek, P. P. (2019). Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values*, *44*(2), 291-314. https://doi.org/10.1177/0162243918793711

Kwon, D. (2024, July 30). *Ai is complicating plagiarism. how should scientists respond?* Nature News. https://www.nature.com/articles/d41586-024-02371-z

Lakim, I., Almazrouei, E., Abualhaol, I., Debbah, M., & Launay, J. (2022, May). A holistic assessment of the carbon footprint of noor, a very large Arabic language model. In *Proceedings of BigScience Episode# 5--Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 84-94). http://dx.doi.org/10.18653/v1/2022.bigscience-1.8

Latour, B. (2004). Which protocol for the new collective experiments? *Experimental cultures*, 17–36.

Laurijssen, S. J., van der Graaf, R., van Dijk, W. B., Schuit, E., Groenwold, R. H., Grobbee, D. E., & de Vries, M. C. (2022). When is it impractical to ask informed consent? A systematic review. *Clinical Trials*, *19*(5), 545-560.

Levine, R. J. (1996). International codes and guidelines for research ethics: a critical appraisal. *The ethics of research involving human subjects: facing the 21st century. Frederick, Maryland: University Publishing Group*, 235-59.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*. https://doi.org/10.48550/arXiv.2211.09110

Liebert, W., & Schmidt, J. C. (2010). Towards a prospective technology assessment: challenges and requirements for technology assessment in the age of technoscience. *Poiesis & Praxis*, *7*, 99-116.

London, A. J. (2021). *For the common good: Philosophical foundations of research ethics*. Oxford University Press.

Long, T. A. (1983). Informed consent and engineering: An essay review. *Business & Professional Ethics Journal*, *3*(1), 59–66.

Lovett, F. (2010). *A general theory of domination and justice*. Oxford University Press.

Lovett, F. (2022). "Republicanism," in The Stanford Encyclopaedia of Philosophy (Fall 2022 Edition), ed. Edward N. Zalta and Uri Nodelman. Accessed at: https://plato.stanford.edu/archives/fall2022/entries/republicanism

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2023). Are Emergent Abilities in Large Language Models Just In-Context Learning? *arXiv preprint arXiv:2309.01809*. https://doi.org/10.48550/arXiv.2309.01809

Lucchi, N. (2023). ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 1-23.

Luna, F. (2009). Elucidating the concept of vulnerability: Layers not labels. *IJFAB: International Journal of Feminist Approaches to Bioethics*, *2*(1), 121–139.

Maas, T., van den Broek, J., & Deuten, J. (2017). Living labs in Nederland: van open testfaciliteit tot levend lab. *Rathenau Instituut*.

MacKay, D. (2018). The ethics of public policy RCTs: The principle of policy equipoise. Bioethics, 32(11),59–67. https://doi.org/10.1111/bioe.12403

MacKay, D., & Chakrabarti, A. (2019). Government policy experiments and informed consent. *Public Health Ethics*, *12*(2), 188–201.

Madiega, T., & Van De Pol, A. L. (2022). Artificial intelligence act and regulatory sandboxes. *European Parliamentary Research Service*, *6*.

Maheshwari, K., & Nyholm, S. (2022). Dominating Risk Impositions. *The Journal of Ethics*, *26*(4), 613–637.

Marres, N. (2020). *What if nothing happens? Street trials of intelligent cars as experiments in participation*. In Techno Science Society (pp. 111–130). Springer, Cham.

Manson, K. (2023, July 5). *The US military is taking generative AI out for a spin*. Bloomberg.com. https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin

Martin, W and Schinzinger, R (1983). *Ethics in Engineering*. New York: McGraw-Hill

McDermott, R., & Hatemi, P. K. (2020). Ethics in field experimentation: A call to establish new standards to protect the public from unwanted manipulation and real harms. *Proceedings of the National Academy of Sciences*, *117*(48), 30014-30021.

McDonald, J., Li, B., Frey, N., Tiwari, D., Gadepally, V., & Samsi, S. (2022). Great power, great responsibility: Recommendations for reducing energy for training language models. *arXiv preprint arXiv:2205.09646*. https://doi.org/10.48550/arXiv.2205.09646

McNeill, P. M. (1993). *The ethics and politics of human experimentation*. CUP Archive.

Melham, K., Moraia, L. B., Mitchell, C., Morrison, M., Teare, H., & Kaye, J. (2014). The evolution of withdrawal: negotiating research relationships in biobanking. *Life sciences, society and policy*, *10*, 1-13.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211.

Meyer, M. N. (2015). Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. Colo. Tech. LJ, 13, 273.

Miller, F. G. (2010). Striking the right balance in research ethics and regulation. *The American Journal of Bioethics*, *10*(8), 65-65.

Ministerie van Algemene Zaken. (2022, June 17). *Impact assessment fundamental rights and algorithms*. Report | Government.nl. Accessed at: https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

Moffatt, B. (2010). Not all human subjects research is exceptional. *The American Journal of Bioethics*, *10*(8), 62–63.

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, *31*(2), 323-327. https://doi.org/10.1007/s11023-021-09557-8

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, *32*(2), 241-268. https://doi.org/10.1007/s11023-021-09577-4

Mollen, J. (2018). *Smart sound sensors will help Dutch Police Nip Street fights (and Weed Farms) in the bud*. The Next Web. https://thenextweb.com/news/1128392

Mollen, J. (2023). Moving out of the Human Vivarium: Live-in Laboratories and the Right to Withdraw. *Journal of Ethics and Emerging Technologies*, *33*(1), 1–22.

Mollen, J. (2024). Towards a Research Ethics of Real-World Experimentation with Emerging Technology. *Journal of Responsible Technology*, 100098.

Mollen, J., Van Der Putten, P., & Darling, K. (2023). Bonding with a Couchsurfing Robot: The Impact of Common Locus on Human-Robot Bonding In-the-Wild. *ACM Transactions on Human-Robot Interaction*, *12*(1), 1-33.

Molnar, P. (2020). Technological testing grounds: Migration management experiments and reflections from the ground up. *EDRI, November Available at edri. org/wp-content/uploads/2020/11/Technological-Testing-Grounds. pdf (accessed 20 April 2021)*.

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, behavior, and social networking*, *16*(9), 708-713.

Morris, T., Manley, D., Northstone, K., & Sabel, C. E. (2017). How do moving and other major life events impact mental health? A longitudinal analysis of UK children. *Health & place*, *46*, 257-266.

Msoroka, M. S., & Amundsen, D. (2018). One size fits not quite all: Universal research ethics with diversity. *Research Ethics*, *14*(3), 1-17.

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869-877.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, DC: Department of Health, Education and Welfare. Accessed March 14th, 2022. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Nelson, R. M., & Merz, J. F. (2002). Voluntariness of consent for research: an empirical and conceptual review. *Medical care*, *40*(9), V-69.

Nevmerzhitskaya, J. (2020). Ethical considerations of Living Labs. *Ethics as a resource. Examples of RDI projects and educational development*.

Nguyen, V. K. (2009). Government-by-exception: Enrolment and experimentality in mass HIV treatment programmes in Africa. *Social Theory & Health*, *7*(3), 196-217.

Nijhawan, L. P., Janodia, M. D., Muddukrishna, B. S., Bhat, K. M., Bairy, K. L., Udupa, N., & Musmade, P. B. (2013). Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research*, *4*(3), 134-140.

Nikolinakos, N. T. (2023). Ethical principles for trustworthy AI. In *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies-The AI Act* (pp. 101-166). Cham: Springer International Publishing.

Nuremberg Code. (1949). *In Trials of War Criminals before Nuremberg Military Tribunals*, Washington, DC: US Government Printing Office.

Nyholm, S. (2023). Minding the Gap (s): Different Kinds of Responsibility Gaps Related to Autonomous Vehicles and How to Fill Them. In *Connected and Automated Vehicles: Integrating Engineering and Ethics* (pp. 1–18). Cham: Springer Nature Switzerland.

OECD (2019). *Digital Innovation: Seizing Policy Opportunities*, OECD Publishing, Paris

Pane, J., Francisca, R. D., Verhamme, K. M., Orozco, M., Viroux, H., Rebollo, I., & Sturkenboom, M. C. (2019). EU postmarket surveillance plans for medical devices. *Pharmacoepidemiology and drug safety*, *28*(9), 1155-1165. https://doi.org/10.1002/pds.4859

Pamplany, A., Gordijn, B., & Brereton, P. (2020). The ethics of geoengineering: A literature review. *Science and Engineering Ethics*, *26*, 3069-3119.

Peterson, M. B. (2013). New technologies should not be treated as social experiments. *Ethics, Policy & Environment*, *16*(3), 349–351.

Petryna, A. (2007). Experimentality: on the global mobility and regulation of human subjects research. *PoLAR: Political and Legal Anthropology Review*, *30*(2), 288–304.

Petryna, A. (2009). When experiments travel: clinical trials and the global search for human subjects.

Pfotenhauer, S., Laurent, B., Papageorgiou, K., & Stilgoe, A. J. (2022). The politics of scaling. *Social Studies of Science*, *52*(1), 3-34.

Phillips, T. (2021). Ethics of field experiments. *Annual Review of Political Science*, *24*, 277–300.

Podschuweit, N. (2021). How ethical challenges of covert observations can be met in practice. Research Ethics, 17(3), 309-327. https://doi.org/10.1177/17470161211008218

Polonioli, A., Ghioni, R., Greco, C., Juneja, P., Tagliabue, J., Watson, D., & Floridi, L. (2023). The Ethics of Online Controlled Experiments (A/B Testing). *Minds and Machines*, 1-27.

Popper, K. (2013). *The poverty of historicism*. Routledge.

Porter, Z., Habli, I., McDermid, J., & Kaas, M. (2024). A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics*, *4*(2), 593-616.

Radder, H. (2009). The philosophy of scientific experimentation: a review. *Automated experimentation*, *1*(1), 1–8.

Ragazzi, F. P. S. M., Kuskonmaz, E., Plájás, I., van de Ven, R. R., & Wagner, B. (2021). *Biometric and behavioural mass surveillance in EU member states: report for the Greens/EFA in the European Parliament* (Leiden University).

Rahman, H. A., Weiss, T., & Karunakaran, A. (2023). The experimental hand: How platform-based experimentation reconfigures worker autonomy. *Academy of Management Journal*, *66*(6), 1803-1830.

Rahman, H. A. (2024, March 10). *Unethical Online Experiments Risk Real World Harm*. Financial Times. https://www.ft.com/content/de6ab765-35e9-4146-8527-e82795062173

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Ranchordas, S. (2021). Experimental regulations for AI: sandboxes for morals and mores. *University of Groningen Faculty of Law Research Paper*, (7).

Raymond, N. (2019). Reboot ethical review for the age of big data. *Nature*, *568*(7752), 277–277.

Raymond, N. (2019). Safeguards for human studies can't cope with big data. *Nature*, *568*(7753), 277–278.

Regnault, J. M. (2003). France's search for nuclear test sites, 1957-1963. *The journal of military history*, *67*(4), 1223-1248.

Renn, O. (2018). Real-world laboratories-the road to transdisciplinary research? *GAIA-Ecological Perspectives for Science and Society*, *27*(1), 1–1.

Resseguier, A., & Ufert, F. (2023). AI research ethics is in its infancy: the EU's AI Act can make it a grown-up. *Research Ethics*, 17470161231220946.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J. (2024). BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. *arXiv preprint arXiv:2411.12990*.

Rheinberger, H. J. (1997). Toward a history of epistemic things: Synthesizing proteins in the test tube.

Rhodes, R. (2010). Rethinking research ethics. *The American Journal of Bioethics*, *10*(10), 19–36.

Richter, E. D., Barach, P., Berman, T., Ben-David, G., & Weinberger, Z. (2001). Extending the boundaries of the Declaration of Helsinki: a case study of an unethical experiment in a non-medical setting. *Journal of Medical Ethics*, *27*(2), 126-129.

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental Science & Technology*, *57*(9), 3464-3466. https://doi.org/10.1021/acs.est.3c01106

Roberts, G. (2011). "PlaceLab studies interaction with Home Tech". *New Bedford Standard-Times*. Retrieved March 3, 2023, from https://eu.southcoasttoday.com/story/news/2004/12/25/placelab-studies-interaction-with-home/50356377007/

Roulet, T. J., Gill, M. J., Stenger, S., & Gill, D. J. (2017). Reconsidering the Value of Covert Research: The Role of Ambiguous Consent in Participant Observation. *Organizational Research Methods*, *20*(3), 487- 517. https://doi.org/10.1177/1094428117698745

Sagarin, E. (1973). The research setting and the right not to be researched. *Social Problems*, *21*(1), 52–64.

Sainz, F. J. (2012). Emerging ethical issues in living labs. *Ramon Llull Journal of Applied Ethics*, (3), 47–62.d

Santoni de Sio, F., Almeida, T., & Van Den Hoven, J. (2024). The future of work: freedom, justice, and capital in the age of artificial intelligence. *Critical Review of International Social and Political Philosophy*, *27*(5), 659-683.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.

Schaefer, G. O., & Wertheimer, A. (2010). The right to withdraw from research. *Kennedy Institute of Ethics Journal*, *20*(4), 329-352.

Schinzinger, R., & Martin, M. W. (1983). Commentary: Informed consent in engineering and medicine. *Business & Professional Ethics Journal*, *3*(1), 67-77.

Schiphol. (2017, February 7). *Test at Amsterdam Airport Schiphol: Quick and easy boarding using facial recognition*. Schiphol Newsroom. https://news.schiphol.com/test-at-amsterdam-airport-schiphol-quick-and-easy-boarding-using-facial-recognition/

Schücklenk, U., & Ashcroft, R. (2000). International research ethics. *Bioethics*, *14*(2), 158-172.

Schwartz, T., Stevens, G., Jakobi, T., Denef, S., Ramirez, L., Wulf, V., & Randall, D. (2015). What people do with consumption feedback: a long-term living lab study of a home energy management system. *Interacting with Computers*, *27*(6), 551-576.

Schwarz, A., & Krohn, W. (2011). Experimenting with the concept of experiment: Probing the epochal break. In *Science transformed? Debating claims of an epochal break* (pp. 119-134). University of Pittsburgh Press.

Seddon, P. (2023, September 29). *AI chatbots do work of civil servants in productivity trial.* BBC News. https://www.bbc.com/news/uk-politics-66810006

Sens, F. (2015). "Lichtproject De-escalate Stratumseind van start." January 9, 2015. https://www.cursor.tue.nl/nieuws/2015/januari/lichtproject-de-escalate-stratumseind-van-start/.

Shuster, E. (1997). Fifty years later: the significance of the Nuremberg Code. *New England Journal of Medicine*, *337*(20), 1436–1440.

Singer-Brodowski, M., Beecroft, R., & Parodi, O. (2018). Learning in real-world laboratories: A systematic impulse for discussion. *GAIA-Ecological Perspectives for Science and Society*, *27*(1), 23–27.

Skinner, E. A. (1985). Action, control judgments, and the structure of control experience. *Psychological Review*, *92*(1), 39.

Skinner, E. A. (1996). A guide to constructs of control. *Journal of personality and social psychology*, *71*(3), 549.

Smith, H. (2021). Clinical AI: opacity, accountability, responsibility and liability. *Ai & Society*, *36*(2), 535-545. https://doi.org/10.1007/s00146-020-01019-6

Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3).

Sommers, R., & Miller, F. G. (2013). Forgoing debriefing in deceptive research: Is it ever ethical? *Ethics & Behavior*, *23*(2), 98-116.

Spicker, P. (2011). Ethical Covert Research. Sociology, 45(1), 118–133. https://doi.org/10.1177/0038038510387195

Spillman, M. A., & Sade, R. M. (2007). Clinical trials of xenotransplantation: waiver of the right to withdraw from a clinical trial should be required. *Journal of Law, Medicine & Ethics*, *35*(2), 265–272.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*. https://doi.org/10.48550/arXiv.2206.04615

Stahl, B. C., Borsella, E., Porcari, A., & Mantovani, E. (2019). Responsible innovation in ICT: Challenges for industry. In *International Handbook on Responsible Innovation* (pp. 367-378). Edward Elgar Publishing.

Stilgoe, J. (2016). Geoengineering as collective experimentation. *Science and Engineering Ethics*, *22*, 851–869.

Stilgoe, J. (2020). Who's driving innovation? *New Technologies and the Collaborative State. Cham, Switzerland: Palgrave Macmillan*.

Susser, D. (2021). Predictive policing and the ethics of pre-emption. *The ethics of policing: New perspectives on law enforcement*, 268–292.

Svensson, S., & Hansson, S. O. (2007). Protecting people in research: a comparison between biomedical and traffic research. *Science and engineering ethics*, *13*, 99-115.

Taylor, L. (2021). Exploitation as innovation: research ethics and the governance of experimentation in the urban living lab. *Regional Studies*, *55*(12), 1902-1912.

Teele, D. L. (2014). Reflections on the ethics of field experiments. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, 115–140.

Traianou, A., & Hammersley, M. (2021). Is there a right not to be researched? Is there a right to do research? Some questions about informed consent and the principle of autonomy. *International Journal of Social Research Methodology*, *24*(4), 443–452.

United States Department of Health and Human Services. "45 CFR 46." Accessed 27 April 2023: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html

Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology*, *35*(4), 88. https://doi.org/10.1007/s13347-022-00577-5

Vallance, S., & McCallum, C. (2023). *Amazon trials humanoid robots to "free up" staff*. BBC News. https://www.bbc.com/news/technology-67163680

Van de Poel, I. (2013). Why new technologies should be conceived as social experiments. *Ethics, Policy & Environment*, *16*(3), 352-355.

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and engineering ethics*, *22*(3), 667-686.

Van de Poel, I. (2017). Society as a laboratory to experiment with new technologies. *Embedding new technologies into society: A regulatory, ethical and societal perspective*, 61-68.

Van de Poel, I. (2017b). Moral experimentation with new technology. In *New perspectives on Technology in Society* (pp. 59-79). Routledge.

Van Heerden, A. C., Pozuelo, J. R., & Kohrt, B. A. (2023). Global mental health services and the impact of artificial intelligence–Powered large language models. *JAMA psychiatry*, *80*(7), 662-664. doi:10.1001/jamapsychiatry.2023.1253

Van Monsjou, D. (2024). "Jumbo Gaat Tests Met AI Doen Voor Tegengaan Winkeldiefstal" n.d. Accessed December 20, 2024. https://tweakers.net/nieuws/218396/jumbo-gaat-tests-met-ai-doen-voor-tegengaan-winkeldiefstal.html.

Van Rijssel, T. I., de Jong, A. J., Santa-Ana-Tellez, Y., Boeckhout, M., Zuidgeest, M. G., van Thiel, G. J., & Trials@ Home Consortium. (2022). Ethics review of decentralized clinical trials (DCTs): results of a mock ethics review. *Drug discovery today*, *27*(10), 103326.

Verbeke, K., Krawczyk, T., Baeyens, D., Piasecki, J., & Borry, P. (2023). Informed Consent and Debriefing When Deceiving Participants: A Systematic Review of Research Ethics Guidelines. *Journal of Empirical Research on Human Research Ethics*, 15562646231173477.

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 941-953).

Waern, A. (2016, May). The ethics of unaware participation in public interventions. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 803-814).

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526-1541. https://doi.org/10.1038/s41562-023-01659-w

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*. https://doi.org/10.48550/arXiv.2206.07682

Weiss, T. (2024, November 12). *Why we are all lab rats in the Digital World*. Nature News. https://www.nature.com/articles/d41586-024-03674-x#author-0

Weinmann, M., Schneider, C., & Brocke, J. V. (2016). Digital nudging. *Business & Information Systems Engineering*, *58*, 433-436.

Whitfield, G. (2019). TRENDS: Toward a separate ethics of political field experiments. *Political Research Quarterly*, *72*(3), 527–538.

Wilson, J & D. Hunter. (2010). *Research Exceptionalism*. The American Journal of Bioethics, 10(8): 45–54.

Wong, R. Y., Boyd, K., Metcalf, J., & Shilton, K. (2020). Beyond checklist approaches to ethics in design. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 511–517).

Wood, M. (2020). *OKCupid plays with love in user experiments*. NY Times. 28 July 2014. <https://www.nytimes.com/2014/07/29/technology/okcupid-publishes-findings-of-user-experiments.html>. Accessed 6 July 2022

World Medical Association. (2013). *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. Accessed 07 January 2022: https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/doh-oct2008/

Zimmermann, V. (2023). Smart cities as a testbed for experimenting with humans?-Applying psychological ethical guidelines to smart city interventions. *Ethics and Information Technology*, *25*(4), 54.

Zhang, J. J. (2016). Research ethics and ethical research: some observations from the

Global South. *Journal of Geography in Higher Education*, *41*(1), 147–154. https://doi.org/10.1080/03098265.2016.1241985

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in generative AI. *arXiv preprint arXiv:2403.02726*.

About the Author

Joost Mollen (1995) was born in Nijmegen, the Netherlands. He completed his Ph.D. on the ethics of real-world technology research at Delft University of Technology between April 2021 and April 2025. Prior to his Ph.D., he obtained his M.Sc. (cum laude) in Media Technology from Leiden University (Netherlands) in 2020 with a focus on human-robot interaction and his B.A. in Media and Culture from the University of Amsterdam in 2017. During his studies, he was a visiting student at KU Leuven's Faculty of Engineering and the University of Edinburgh's Faculty of Humanities.

List of Publications

Mollen, J. (2025). LLMs beyond the lab: the ethics and epistemics of real-world AI research. *Ethics and Information Technology*, 27(1), 1–11.

Kudina, O., **Mollen, J**.*, Viader Guerrero, J., Muravyov, D., & Bermudez, J. P. (2025). Teaching and Crafting Human-Robot Relational Ethics. *SEFI Journal of Engineering Education Advancement*, *2*(2), 6–31. (*shared first authorship)

Mollen, J. (2024). Towards a Research Ethics of Real-World Experimentation with Emerging Technology. *Journal of Responsible Technology*, 100098.

Mollen, J. (2023). Moving out of the Human Vivarium: Live-in Laboratories and the Right to Withdraw. *Journal of Ethics and Emerging Technologies*, 33(1), 1.

Mollen, J., Van Der Putten, P., & Darling, K. (2023). Bonding with a Couchsurfing robot: The impact of a common locus on human-robot bonding in the wild. *ACM Transactions on Human-Robot Interaction*, 12(1), 1-33.

Book reviews and commentaries

Mollen, J. (2023). A. Fejerskov, The Global Lab: Inequality, Technology, and the Experimental Movement. *Prometheus*, 39(3), 189-194.

Chen, S. S., Brenna, C. T., **Mollen, J**., & Das, S. (2025). Which Risks Can Undermine Benefits in Research?. *The American Journal of Bioethics*, *25*(5), 86-88.

The Simon Stevin Series in Ethics of Technology is an initiative of the 4TU Centre for Ethics and Technology. 4TU.Ethics is a collaboration between Delft University of Technology, Eindhoven University of Technology, University of Twente, and Wageningen University & Research. Contact: info@ethicsandtechnology.eu

Books and Dissertations

Volume 1: Lotte Asveld, '*Respect for Autonomy and Technology Risks'*, 2008

Volume 2: Mechteld-Hanna Derksen, *'Engineering Flesh, Towards Professional Responsibility for 'Lived Bodies' in Tissue Engineering',* 2008

Volume 3: Govert Valkenburg, *'Politics by All Means. An Enquiry into Technological Liberalism'*, 2009

Volume 4: Noëmi Manders-Huits, *'Designing for Moral Identity in Information Technology'*, 2010

Volume 5: Behnam Taebi, *'Nuclear Power and Justice between Generations. A Moral Analysis of Fuel Cycles'*, 2010

Volume 6: Daan Schuurbiers, *'Social Responsibility in Research Practice. Engaging Applied Scientists with the Socio-Ethical Context of their Work'*, 2010

Volume 7: Neelke Doorn, *'Moral Responsibility in R&D Networks. A Procedural Approach to Distributing Responsibilities'*, 2011

Volume 8: Ilse Oosterlaken, *'Taking a Capability Approach to Technology and Its Design. A Philosophical Exploration'*, 2013

Volume 9: Christine van Burken, *'Moral Decision Making in Network Enabled Operations'*, 2014

Volume 10: Faridun F. Sattarov, *'Technology and Power in a Globalising World, A Political Philosophical Analysis'*, 2015

Volume 11: Gwendolyn Bax, *'Safety in large-scale Socio-technological systems. Insights gained from a series of military system studies'*, 2016

Volume 12: Zoë Houda Robaey, *'Seeding Moral Responsibility in Ownership. How to Deal with Uncertain Risks of GMOs'*, 2016

Volume 13: Shannon Lydia Spruit, *'Managing the uncertain risks of nanoparticles. Aligning responsibility and relationships'*, 2017

Volume 14: Jan Peter Bergen, *Reflections on the Reversibility of Nuclear Energy Technologies*, 2017

Volume 15: Jilles Smids, *Persuasive Technology, Allocation of Control, and Mobility: An Ethical Analysis*, 2018

Volume 16: Taylor William Stone, *Designing for Darkness: Urban Nighttime Lighting and Environmental Values*, 2019

Volume 17: Cornelis Antonie Zweistra, *Closing the Empathy Gap: Technology, Ethics, and the Other*, 2019

Volume 18: Ching Hung, *Design for Green: Ethics and Politics for Behavior-Steering Technology*, 2019

Volume 19: Marjolein Lanzing, *The Transparent Self: a Normative Investigation of Changing Selves and Relationships in the Age of the Quantified Self*, 2019

Volume 20: Koen Bruynseels, *Responsible Innovation in Data-Driven Biotechnology*, 2021

Volume 21: Naomi Jacobs, *Values and Capabilities: Ethics by Design for Vulnerable People*, 2021

Volume 22: Melis Baş, *Technological Mediation of Politics. An Arendtian Critique of Political Philosophy of Technology*, 2022

Volume 23: Mandi Astola, *Collective Virtues. A Response to Mandevillian Morality*, 2022

Volume 24: Karolina Kudlek, *The Ethical Analysis of Moral Bioenhancement. Theoretical and Normative Perspectives*, 2022

Volume 25: Chirag Arora, *Responsibilities in a Datafied Health Environment*, 2022

Volume 26: Agata Gurzawska, *Responsible Innovation in Business. A Framework and Strategic Proposal*, 2023

Volume 27: Rosalie Anne Waelen, *The Power of Computer Vision. A Critical Analysis*, 2023

Volume 28: José Carlos Cañizares Gaztelu, *Normativity and Justice in Resilience Strategies*, 2023

Volume 29: Martijn Wiarda, *Responsible Innovation for Wicked Societal Challenges: An Exploration of Strengths and Limitations*, 2023

Volume 30: Leon Walter Sebastian Rossmaier, *mHealth Apps and Structural Injustice*, 2024

Volume 31: Haleh Asgarinia, *Privacy and Machine Learning-Based Artificial Intelligence: Philosophical, Legal, and Technical Investigations*, 2024

Volume 32: Caroline Bollen, *Empathy 2.0: What it means to be empathetic in a diverse and digital world*, 2024

Volume 33: Iris Loosman, *Rethinking Informed Consent in mHealth*, 2024

Volume 34: Benjamin Hofbauer, *Governing Prometheus. Ethical Reflections On Risk & Uncertainty In Solar Climate Engineering Research*, 2024

Volume 35: Madelaine Ley, *It's not (just) about the robots: care and carelessness across an automated supply chain*, 2024

Volume 36: Arthur Gwagwa, *Re-imagining African Unity in a Digitally Interdependent World*, 2024

Volume 37: Jonne Maas, *Freedom in the Digital Age: Designing for Non-Domination*, 2025

Volume 38: Nynke van Uffelen, *Reconceptualising Energy Justice in light of Normative Uncertainties*, 2025

Volume 39: Cindy Friedman, *The Ethics of Humanoid Robots*, 2025

Volume 40: Tom Hannes, *What do We Call the World? A Plea for Developing an Anthropocene Morality Based on a Non-Axial Rereading of Buddhism*, 2025

Volume 41: Joost Mollen, *Prototype Ethics. Foundations for the Research Ethics of Real-World Technology Research*, 2025

Simon Stevin (1548-1620)

'Wonder en is gheen Wonder'

This series in the philosophy and ethics of technology is named after the Dutch / Flemish natural philosopher, scientist, and engineer Simon Stevin. He was an extraordinarily versatile person. He published, among other things, on arithmetic, accounting, geometry, mechanics, hydrostatics, astronomy, theory of measurement, civil engineering, the theory of music, and civil citizenship. He wrote the very first treatise on logic in Dutch, which he considered to be a superior language for scientific purposes. The relation between theory and practice is a main topic in his work. In addition to his theoretical publications, he held a large number of patents, and was actively involved as an engineer in the building of windmills, harbours, and fortifications for the Dutch prince Maurits. He is famous for having constructed large sailing carriages.

Little is known about his personal life. He was probably born in 1548 in Bruges (Flanders) and went to Leiden in 1581, where he took up his studies at the university two years later. His work was published between 1581 and 1617. He was an early defender of the Copernican worldview. He died in 1620, but the exact date and the place of his burial are unknown. Philosophically, he was a pragmatic rationalist. For him, wonder about a phenomenon, however mysterious, should be the starting point for seeking understanding or even ultimate explanation through human reasoning. Hence the dictum 'Wonder is no Wonder' that he used on the cover of several of his books.

To make technologies that work, they need to be exposed to complex 'real-world' environments to evaluate their performance or impact. Examples include tests with self-driving cars on public roads, predictive policing technologies on nightlife streets, or mood-altering algorithms on social media platforms. As this 'real-world research' has become widespread, scholars have drawn attention to its ethical concerns and the absence of research ethics governance, such as ethics guidelines and independent oversight. However, this scholarly attention is fragmented across different disciplines, and it is unclear to what extent existing research ethics principles and norms can capture the common ethical challenges of real-world research. This thesis addresses these gaps. It argues that real-world research shares common ethical salient characteristics, such as 'coupling' options to subjects, and that real-world research needs research ethics governance, but that the content of this research ethics governance cannot be wholly based on existing research ethics principles and norms. This is because real-world research raises novel ethical challenges, and many existing research ethics norms, such as informed consent or the right to withdraw, cannot be upheld without severely altering the practice. Thus, through a comprehensive analysis, this thesis contributes to the groundwork for a new research ethics of real-world technology research.

4TU. Centre for Ethics and Technology

TU Delft Technische Universiteit Delft

TU/e Eindhoven University of Technology

University of Twente.

Wageningen University & Research