A Framework for Systematically Applying Humanistic Ethics when Using AI as a Design Material

KYLE DENT
RICHELLE DUMOND
MIKE KUNIAVSKY
Palo Alto Research Center (PARC)

KEYWORDS
Ethics, Technology and Society, AI Design.

HOW TO CITE
Dent, Kyle, Richelle Dumond and Mike Kuniavsky. 2019.
“A Framework for Systematically Applying Humanistic
Ethics when Using AI as a Design Material”. *Temes de
Disseny* 35: 178-197.

As machine learning and AI systems gain greater capabilities and are deployed more widely, we —as designers, developers, and researchers— must consider both the positive and negative implications of their use. In light of this, PARC’s researchers recognize the need to be vigilant against the potential for harm caused by artificial intelligence through intentional or inadvertent discrimination, unjust treatment, or physical danger that might occur against individuals or groups of people. Because AI-supported and autonomous decision making has the potential for widespread negative personal, social, and environmental effects, we aim to take a proactive stance to uphold human rights, respect individuals’ privacy, protect personal data, and enable freedom of expression and equality.

Technology is not inherently neutral and reflects decisions and trade-offs made by the designers, researchers, and engineers developing it and using it in their work. Datasets often reflect historical biases. AI technologies that hire people, evaluate their job performance, deliver their healthcare, and mete out penalties are obvious examples of possible areas for systematic algorithmic errors that result in unfair or unjust treatment. Because nearly all technology includes trade-offs and embodies the values and judgments of the people creating it, it is imperative that researchers are aware of the value judgments they make and are transparent about them with all stakeholders involved.

INTRODUCTION

1. AI brings many challenges

Every technology, from fire to nuclear power to social networks, brings unexpected consequences with it, and the greater the power of the technology, the more serious the consequences. It’s impossible to entirely avoid the unexpected consequences of a new technology, but we believe it’s possible to create practices that can, to some extent, anticipate and mitigate consequences that have a near-term and direct negative impact on the people interacting with those technologies.

Artificial intelligence (in all its forms, from simple pattern recognition to complex planning, prediction, and reasoning) is incredibly powerful, so the unexpected consequences of its effects are particularly concerning. Our experience as designers of AI algorithms, user experiences, and AI-driven business models has shown us first-hand the challenges of working with this new design material and how important it is to create ethical design practices that simultaneously warn designers of dangerous uses of AI and give them tools with which to mitigate potential uses of AI.

Even listing the challenges that AI poses is itself daunting, since anything that computers touch (which is

nearly everything) becomes a site of potential negative consequences when AI is involved. Some of the challenges it poses, such as the trolley problem for autonomous driving or face recognition technology trained with biased data sets, are clear. Others, such as the impact of automation on jobs, the algorithmic exploitation of cognitive biases, and coercive use by oppressive companies and governments, are emerging. Most of the challenges AI poses are still completely unforeseen, and we probably won’t know about them until the technologies involved have already been broadly deployed.

In this paper we do not intend to provide solutions to the current challenges posed by AI. Instead we provide an ethical review framework for designers participating in the development of AI-driven experiences to use in evaluating potentially negative implications of their work and allowing them to determine for themselves, and highlight for the organization for which they work, specific dangers of their own work. We describe the efforts of our work at the Palo Alto Research Center (formerly Xerox PARC). PARC has a long history in Silicon Valley and has been at the cutting edge of many innovations. As we see more real-world effects from our work and the work of others in AI, we want to be sure that we understand the potential impact and think through ways to mitigate negative consequences.

Among our highest priorities is that when there are specific risks or assumptions, they need to be surfaced and communicated to all stakeholders, including those affected by the AI. Technology development is a continual negotiation between the creators of a technology, the organizations who sponsor that work, and end users who are the consumers of, or targets of, that technology. There is not just a single ethical decision to make in the development of a technology (other than an explicit decision to make something intended to hurt people, damage the environment, or undermine social order – in which case there is no process and one should be loath to consider working on such a project). Instead, there is a series of decisions about the focus, scope, transparency, and experience of a technology that appears throughout a project that require ethical consideration, deliberation, discussion, and consensus.

2. We think that emerging protocols and checklists are insufficient

Google’s AI principles, to cite one example, are anemic (Google LLC 2018). Broad philosophical discussions of the trolley problem, in all its various forms, only get at one small sliver of one kind of negative impact a single technology has. Talking about it is like talking about the negative impact of huffing gasoline without discussing all of the other negative effects gasoline has, from its toxicity, to its flammability, to the environmental impact of internal combustion engines. Yes: people shouldn’t huff gasoline, but if that’s what you talk about, you’re missing the big picture and many aspects of it that really matter.

In general, simple checklists don’t go far enough. It’s also not just a question of standards, but processes: design review process, user research process, and data collection process. Processes must include oversight by groups with diverse backgrounds as well as transparency of data collection and algorithms. Design must include the ability to interrogate systems with clear mechanisms to control them. Processes should incorporate warning signals as procedural fuses that highlight when development is headed into dangerous areas and specify what happens if a process has spiraled out of control.

3. Design process governance

We believe that, like medical research, chemical plant processes, highway driving, and many other everyday practices, there needs to be an AI design governance model. Projects must have the ability to understand and shape participants such as data scientists, social scientists, organizational and non-organizational policy makers, and designers of all sorts. The goal is not to slow down development with additional bureaucracy but to have thought about what kinds of things may get broken when you’re moving fast, what the plan is to identify when they’re breaking before they’ve broken completely, and how to fix them before they’ve broken irreparably. Our process borrows from other similar approaches like Institutional Review Boards (IRBs), participatory design, design ethics, and Computer Professionals for Social Responsibility (CPSR)

There is an active, ongoing conversation about the ethical implications of AI both in the media and in academia with several recent conferences and workshops organized around themes related to AI for social good. The different perspectives at these gatherings tend to consider ethics and AI in one or two ways: either as a technical challenge (the problem of incorporating moral decision making into intelligent agents) or ethics as a way to rein in a disruptive and powerful societal force. These are worthy discussions indeed but, apart from the (very real) problem of bias in data (O'Neill 2016), there is very little discussion about the current and immediate harm from intelligent systems and how it should be practically addressed today. At the AAAI 2018 Spring Symposium on AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents, for example, of sixteen accepted papers, fourteen covered ways to include moral decision making within agents and only two (Williams 2018; Dent 2018) discussed the current ethical impact of AI systems on human lives. As a technical conference, this may not be surprising. But nearly all of the other papers were necessarily based on an imagined future state where highly sophisticated AI agents are fully participating within human spheres. The community is currently deploying much less sophisticated AI technologies that have harmful effects on people and society; however, for the most part, developers’ priorities are elsewhere. We have not only been lacking in our responsibility to imagine all the ways AI could impact the world around us, we have caused harm by not being critical of the risks that our AI systems can inflict.

Another activity related to AI and ethics is the creation of guidelines. Several academic and industry guidelines have recently been published. Even here, as Hagendorff (2019) points out “...very little to nothing has been written about the tangible implementation of ethical goals and values.” Hagendorff reviews fifteen different sets of guidelines. He makes it clear that there is no evidence of tangible impact from guidelines, pointing out that one difficulty is that they are composed of language like *justice*, *transparency*, and *human-centered*, without definitions or any precision about what the terms mean. Moreover, voluntary guidelines have the inherent problem of lacking enforcement mechanisms.

A study published last year (McNamara, Smith and Murphy-Hill 2018) considered the question of whether or not the ethics guidelines published by the Association for Computing Machinery has any influence on software-related ethical decisions. Participants were given a series of relevant vignettes and then asked questions about choices they would make given the circumstances. The researchers found no statistically significant difference between responses from the study group instructed to take the guidelines into consideration and a control group which did not see them.

Even so, we believe that publishing guidelines helps to bring issues to society’s attention and informs the debate. It’s helpful to state a set of principles in the context of AI that most would agree will produce better outcomes for humanity. One of the better guideline examples, AI4People (Floridi et

al. 2018) advocates for attitudes most likely to produce beneficial uses of AI. We do not expect to see concrete changes in designers’ and researchers’ behavior and choices because of the existence of guidelines, however.

Some of the more practical guidelines unrelated to AI from within industry are instructive. Microsoft has produced a toolkit for designers with considerations for building inclusiveness into their products. Interestingly, although Microsoft focuses on accessibility, the toolkit originated in AI. When working on Cortana, Microsoft’s AI assistant, the designers faced challenges posed by the limitations of conversational user interfaces.

- 1. Microsoft Inclusive Design Toolkit*
- Microsoft describes inclusive design as “a design methodology that enables and draws on the full range of human diversity. Most importantly, this means including and learning from people with a range of perspectives.” (Microsoft Corporation n.d.)
- Microsoft promotes inclusive design by focusing on three core principles:
1. *Recognize exclusion:* Inclusive design reminds us that our personal experiences and abilities create a bias. When we don’t account for other experiences, and base design solutions on ourselves, we exclude others who have different personal experiences and abilities.
 2. *Learn from diversity:* When a mismatch between a person’s ability and a situation occurs, people adapt. These adaptations provide insights into ways that services can be improved.
 3. *Solve for one, extend to many:* Designing for disabilities also benefits other users who are experiencing a situational limitation. The toolkit provides an example of a device designed for a person who has one arm that also aids in the tasks of a person with a temporary wrist injury or a new parent holding an infant.
- Microsoft Inclusive shows that inclusive design should not be seen as a taxation after the design process has completed but rather should be part of the design process. We believe the same is true for ethics as part of the AI design process, i.e. AI ethics as a design material.

2. UX research

Having a diverse team, covering more than race and sex, includes diversity in gender identity, ability status, socio-economic status, education level, religion, country of origin, age, technical literacy, and so on. Such diversity in teams provides easy access to a range of experiences. A group with all members sharing similar life experiences will never achieve the same result as one with a variety of perspectives to draw from. Diverse teams, where all voices can be heard, gain great insight into different worldviews with minimal additional effort.

The meager diversity of technical teams in Silicon Valley has, without a doubt, contributed to the harm AI has caused over the past few years. Even with a diverse team, building empathy based on just your team’s experience only gets you so far. Beyond the indisputable need for more di-

versity on technical teams, a case can be made for more extensive UX research as well.

Often, when designers think about user-experience research, we have our clients in mind. Designers and engineers need a fundamental shift to include the experience of all stakeholders during the design process. The UX design of an AI project should not be focused just on the usability of an interface and whether or not the UI makes sense to whoever is interacting with it (e.g. if a “submit” button should be aligned to the left or right of the screen). The design of AI needs to include the experiences of the people it makes decisions about. For example, it needs to consider the experience of the person that is being denied probation based on machine learning models.

Mike Monteiro, designer at Mule Design, released his designer code of ethics in 2017 in the form of a red booklet titled “Design Ethics.” Monteiro argues, among other topics, that there are no such things as edge cases when designing. He suggests that by claiming certain groups represent an edge case, we are inherently marginalizing them (Monteiro 2017). Monteiro uses Facebook as an example. With Facebook claiming it now has over two billion users, 1% (a number that might be considered an edge case) means 20 million people. A decision that ignores a 1% margin excludes and marginalizes 20 million people. As Monteiro reminds us, “[t]hey are not edge cases. They are human beings and we owe them our best work.” (Monteiro 2017)

The severity of the impact matters. Netflix recommending a terrible movie because you secretly watched *The Great British Bake Off* is annoying but soon forgotten after browsing an additional few minutes. Being hit by an autonomous vehicle because it misidentified you, has deadly consequences (Marshall and Davies 2018). The stakes are high when it comes to automation.

TOWARDS AN ETHICAL FRAMEWORK
AND PROCESS IN AI DESIGN

When we, as AI designers and researchers, reason about our work and its impact on the world, certain methods for focusing attention and thinking critically about potential outcomes are better than others. Better methods are more likely to produce truthful, credible, and more beneficial results. The difference between a cursory consideration and a thorough process can have significant consequences. With this in mind, PARC researchers have developed a design methodology for working with the potentially dangerous material of AI.

Ethical frameworks with organizational enforceability such as we propose are not without precedent. One that most closely resembles and inspired our approach is described in the Belmont Report (Belmont 1979). Created by the U.S. government, the report discusses and presents ethical principles and guidelines for the protection of human subjects of research studies. There were several well-known incidents that led to the development of the report, primarily the Tuskegee Syphilis Study, which was run from 1932 until 1972 by the U.S. Public Health Service. The study is now

Fig. 1. When designing and deploying AI systems, decisions made about the obvious stakeholder will impact many others around them. All of those affected are just as much stakeholders in the outcome of a design as the designer, the client, and what has traditionally been called the "user." When designing AI systems, it's important to include the experiences of all the people affected. Illustration: Ying Xiao

notorious for its violation of ethical standards as researchers knowingly withheld treatment from sick patients.

It is a tragic irony that, most likely, the Tuskegee researchers and others like them convinced themselves they were doing good. For those who might have been less assured, a brief and misguided moral calculation concluding with some net benefit to society no doubt satisfied any concerns. AI researchers and designers now find themselves in a similar situation. Tremendous advantages of AI-powered technology are covered regularly in media reports. Less visible are the negative consequences that have already been seen with very little thought about the potential of other harms. In the current AI design culture, even well-intentioned practitioners demonstrate an inability to understand their own responsibility in developing new applications of AI, even when it's making decisions about human lives or interacting with humans in ways that can risk death (Wakabayashi 2018).

Across history many other documents have set out to establish rules to govern conduct to minimize harm from human interactions, especially in the face of asynchronous power relationships (government over individuals, advantaged over disadvantaged, privileged over disenfranchised, expert over unskilled, etc.). Some of these now have the force of international law behind them. Regrettably, documents like the Nuremberg Code, the Universal Declaration of Human Rights, and the Declaration of Helsinki were attempts to restore moral order following episodes of human rights disasters. Across these documents, we discovered a few high-level principles relevant to our work that has guided our thinking:

1. *Respect for individuals*
2. *Safety and security of person*
3. *Respect for individuals' right to privacy, protection of data, freedom of expression and participation in cultural life*
4. *Equality and fairness*
5. *Benefits to society must not come at the cost of others*

AI DESIGN GUIDELINES

Given our location in the heart of Silicon Valley in the United States, we necessarily have a certain perspective on issues of technology and society. We are steeped in the latest innovations and have the chance to observe their effects, both positive and negative, on populations and subpopulations of people. Our approach when considering the potential effects from new technology is not to consult a list of simple ethical standards hoping to be nudged in the right direction. Real-world ethics are rarely simple. Ethical excellence begins by facing hard questions leading to complex discussions that deal with difficult trade-offs. Moreover, questions and discussions must place the ethical considerations in the context of the expected use of a particular technology. We propose a committee of relative experts within the organization who can help researchers and designers negotiate the ethical landscapes of their projects. Committee members continue their own ethical education and refine their expertise through training, topical readings with discussion, and

ongoing discussions about ethically relevant incidents and events in the world.

Below, a set of guiding questions is given that a committee can use to prompt project discussions to drive the conversation towards the complexity, not away from it.

1. Preliminary Checklist

Respect for designers' time and attention requires that ethical reviews are not overly cumbersome and aren't demanded when not required. A simple checklist helps to assess the potential for harm and the degree of impact.

When the AI learns from existing data, ask yourself:

- Does your project data represent individuals or populations of people?
- Yes/No
- Does the data contain individual personal attributes (especially protected attributes such as race, sex, gender identity, ability, status, socio-economic status, education level, religion, country of origin)?
- Yes/No
- Is the goal to make predictions about people's behavior?
- Yes/No
- Is the goal to classify people or otherwise make predictions about them?
- Yes/No
- Is the goal to make decisions about people or populations of people that could have a significant impact on their lives (e.g. job performance, judicial sentencing, fraudulent activity, etc.)
- Yes/No

When AI participates in physical space as with cameras, microphones, sensors, or anything that records human likenesses or activity:

- Is it in a public place?
- Yes/No
- Is it hidden from anyone who might be recorded?
- Yes/No
- In other words, could subjects be recorded without knowing they are being recorded?
- Yes/No
- Is it possible that any members of vulnerable populations (this could be any disadvantaged sub-segment of an overall population, e.g. children, prisoners, refugees, people facing discrimination) might be recorded?
- Yes/No
- Is there heavy equipment or is it operating at high speed?
- Yes/No

Answering 'yes' to any of the above questions indicates additional scrutiny is most likely needed.

2. Comprehensive Guidelines

When a project requires more careful consideration, we propose a comprehensive set of questions to provoke discussion and exploration for potential harm. Not all of them are relevant to all projects. They are designed to prompt ideas that committee members will use to formulate questions related to the project under consideration. The subsequent discussions are meant to lead designers to consider many aspects of the technology. We organize the questions along the dimensions of human data, social and environmental impact, physical interaction, malicious intent, and ecological impact. Some of these questions have been adapted from the FAT/ML organization's (Fairness,

Fig. 2. Processes must include oversight by groups with diverse backgrounds as well as transparency regarding data collection and algorithms. The design must include the ability to interrogate systems with clear mechanisms to control them. Transparency continues after deployment as well, with regular feedback about real-world performance and the operation of the system. Illustration: Ying Xiao

2.1. Human Data

Respect for people as individuals is essential. Data collection of personal information and behavior may include less than fully informed consent from the subjects. Think about how the analysis might affect each subject involved. If the data represents any vulnerable individuals (the elderly, e.g.) or individuals with diminished autonomy (children, e.g.), special consideration should be given to their protection. People should have their individuality and autonomy respected, and steps should be taken to protect them from harm.

› Have you taken appropriate data protection steps given the sensitivity of the data? Any personally identifiable information is potentially sensitive and, in many contexts, such as medical information, requires different levels of handling when managing data about people.

› If the project is for a public entity, can you disclose the sources of your data? Where the data came from, who collected it, and for what purposes can affect not only the quality of the data but may also inform other qualities including possible biases.

› Have you confirmed that your data accurately reflects the real-world situation for the problem you are trying to solve? Can you also consider alternative data sources? Getting the right data can be expensive and sometimes even impossible. Working with data because it's what's available is a surprisingly common practice. It has also led to the unfair treatment of individuals due to mistaken assumptions and wrong conclusions.

› Have you checked the internal consistency of the data (through random sampling, for example)? Should you report any issues to stakeholders? How will you assess the data for both explicit and implicit biases? Consider specific subpopulations too. Perform appropriate checks for statistical biases and think through whether or not the data matches real-life phenomena. It is important to communicate all limitations and assumptions to all stakeholders.

› Will the benefits of your design extend to the entire population or could there be subgroups who are inadvertently excluded? In instances where there is a known risk of discrimination, include details of the application and its functions along with samples of the data and details about the data source. Is there a process to take effective action to mitigate discrimination?

› For projects that have a significant risk of causing human rights abuse, is it possible to submit them for independent third-party audits? Third-party involvement may not always be appropriate since many of us do work for private companies; however, if possible, an independent review of the system and its outcomes when it's deployed is a good check on your work especially if the system is making decisions that could have a significant impact on people's lives.

› Can stakeholders audit the system? At a minimum, you should

not be delivering black-box systems that are not inspectable and subject to verification by those involved.

› Is there a plan to communicate the decisions about trade-offs, project assumptions, shortcomings, error rates, etc. to stakeholders? Throughout the design and development process, you have had to make decisions about assumptions and trade-offs. If those are not captured and communicated, they will never be known beyond the development team.

› Is there a process for people to correct data or contest erroneous decisions? Systems are fallible and never operate with 100% accuracy. People affected by them must have a way to question the results.

› If appropriate, is there a way for data to be removed when necessary (e.g., the right to be forgotten as practiced in the EU and Argentina)?

2.2. Social impact

AI is increasingly becoming a more prominent part of many aspects of our lives, including decision-making processes that could alter their course. Successful designers view possible social impact not as an afterthought or hindrance to a project but actively think about how to prevent harm while the system is being developed.

2.2.1. AI decisions that could bias or alter societal norms

› Are there any subgroups who benefit more or less from the system? Give special consideration to the effects of errors in the system. Are any groups likely to be harmed more significantly than others?

› Think about what services, products, and industries – and perhaps jobs – might be replaced by the deployment of the algorithm/system. How will this impact society? You may not be able to do anything about it, and it might still be appropriate to develop the innovation for the long-term, but from an insider's point-of-view, you may be able to offer insights or suggestions that could mitigate the negative effects your system will have on communities or other systems and groups.

› When using human data, do the benefits outweigh the risks to those involved? Consider the benefits of a project and the potential harm to those affected by it. For example, researchers at Stanford created a vision application that predicts sexual preference by analyzing images of faces. There are almost no benefits from the technology but great potential costs when it comes to impartiality and fairness. As The Economist put it, "[i]n parts of the world where being gay is socially unacceptable, or illegal, such an algorithm could pose a serious threat to safety." (Economist 2017)

› Are the decisions produced by an algorithmic system explainable to the people affected by those decisions? Explanations help with the validation of results and build trust in the system.

› Useful exercise: Imagine yourself subject to all possible outcomes of a system, are they all equally fair and just from all points of view? (Rawls 1999). As you design a system, consider yourself in a position where you

are subject to its decisions and predictions. We cannot eliminate all of our inherent biases, but imagine yourself with different natural abilities and a different position in society with different identities related to gender, race, and nationality.

2.3. Environmental impact

Resource consumption of a deployed system itself might be an overlooked aspect of the design.

› Consider the energy impact of the computational resources necessary for the project. Is there a way to minimize that impact? Can the computation be handled differently to use less energy? Do the resources provisioned for the project match the requirements, or are they too excessive?

› Does the algorithm/system/design encourage unsustainable behavior? The innovation of coffee pods, for example, has made single-cup, on-demand cups of coffee tremendously convenient, but the pods are generally not recyclable and end up in landfills.

2.4. Physical interaction

Much of the concern related to AI has to do with respecting individuals' privacy and rights. Risk of physical injury may also be a concern. Broadly speaking, users should be aware of possible risks and developers must think through potential unintended consequences of interacting with the technology.

› Is there the equivalent of informed consent for the usage? Informed consent means that users fully understand the facts and consequences of their participation. Simple disclaimers and dense usage agreements are the opposite of informed consent.

› Will individuals interacting with the technology understand the capabilities and limits of it? In May 2016, there was a widely publicized crash involving a Tesla Motors car being driven in computer-assisted mode (Habib 2017). It appears the driver had undue faith in the capabilities of the car. The driver used computer-assisted mode in conditions it was expressly not designed for, resulting in his death.

› Does the design minimize potential risk?

› Do the warning signals match the severity of the danger?

2.5. Misuse and malicious intent

An often-overlooked concern is how a technology might be misused. Misuse can happen by mistake, overconfidence in the technology, or by bad actors trying to subvert it. In all cases care should be taken to mitigate potential harm and provide appropriate restrictions to limit misuse.

› Are there guardrails in place to prevent off-label usage (the intentional or accidental misuse of the design)? The Tesla accident serves as a good example here too. Tesla might have been able to design the controls in such a way that drivers would not be able to use them where it is unsafe to do so.

› Has the system been adequately secured to prevent manipulation from outside? If physical devices are used, are they adequately secured from hacking or subversion? For example, in 2016 IoT devices including baby

monitors were hijacked to launch a massive infrastructure attack against large sections of the internet (Sanger and Perlroth 2016).

2.6. Post deployment

The responsibility of a design or system doesn't end at handoff or deployment. Although every effort should be made to prevent unexpected outcomes, considering a sunset plan before delivering the project to the client can prevent confusion about how to solve unforeseen consequences.

› Do individuals and groups have access to meaningful remedy and redress? This may include, for example, creating clear processes for redress following adverse decisions or effects.

› What will the reporting process and process for recourse be?

› Is there a plan for what to do if the project has unintended consequences? This may be part of a maintenance plan and should involve post-launch monitoring plans.

CONCLUSION

Through these principles and guidelines, we aim to include the discussion around AI Ethics as a natural component of AI projects designed and engineered by PARC. We want our process of considering ethics and mitigating potential risks to be routine for the teams working with AI. We believe that with the experience of discussing possible outcomes, extensive understanding of the experiences of all stakeholders involved, and creating a space to expose our biases to scrutiny we nurture an innate competency of ethical design and engineering.

BIOGRAPHIES

Kyle Dent
Palo Alto Research Center

Kyle Dent is an AI researcher and data scientist interested in the interplay between people and technology. He leads R&D projects as manager of the CHAT group and co-chairs the AI Ethics Review Committee. He specializes in technology and society, ethical considerations of AI and machine learning, conversational agents and natural language processing, text analysis, information extraction, data science, and complex systems.

Richelle Dumond
Palo Alto Research Center

Richelle Dumond is a UX Designer and Researcher in the Innovation Services Group. She works on a variety of projects ranging from healthcare solutions to the future of work, combining the novel technologies invented at PARC with human-centered insights. She co-chairs PARC's AI Ethics Review Committee that is the recipient of an AIJ grant promoting AI Research. Before joining PARC, Richelle worked for Simon Dogger, a designer whose loss of sight hasn't stopped him from creating. As a volunteer at the Bureau of Fearless Ideas in Seattle, Richelle helped young humans to develop their creative and expository writing skills and staffed the Space Travel Supply Company. Richelle received the Louis G. Marsh Top Scholar Award in Design with which she earned her Master of Design in Interaction Design from the University of Washington. She completed a Bachelor of Design in Communication Design at the Design Academy Eindhoven in her native Netherlands.

Mike Kuniavsky
Palo Alto Research Center

Mike Kuniavsky is a user experience designer, researcher, and author. A twenty-year veteran of digital product development, Mike designs products, business processes, and services at the leading edge of technological change.

Prior to joining PARC, he co-founded several successful user experience centered companies, including ThingM, which designs and manufactures ubiquitous computing and Internet of Things products, and Adaptive Path, a well-known design consultancy. He specializes in multi-device interactions, cloud-based service design, and design of hardware products connected to cloud-based services. His background includes design for social analytics, consumer electronics, appliances, image retrieval, RGB LEDs, and financial services.

He has worked with some of the world's top technology companies, such as Samsung, Sony, Nokia, Whirlpool, and Qualcomm, to design new products, guide product strategy, and create user-centered design and development cultures.

He is the author of "Observing the User Experience: A Practitioner's Guide to User Research" and "Smart Things: Ubiquitous Computing User Experience Design" both of which are used as standard university textbooks. He received a dual major B.S./B.A. in Computer Science and Film/Video Studies from the University of Michigan.

<div></div>	<div></div>
REFERENCES	
<div></div>	

Belmont. 1979. *Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. accessed December 20, 2018, http://hhs.gov/ohrp/humansubjects/guidance/belmont.html, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Dent, Kyle. 2018. "Ethical Considerations for AI Researchers." *AAAI 2018 Spring Symposium on AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents* . Stanford, CA: Association for the Advancement of Artificial Intelligence.

Economist. 2017. "Keeping a straight face." *The Economist*, September 9: 72-73.

FAT/ML. n.d. *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*. Accessed December 20, 2018. http://www.fatml.org/resources/principles-for-accountable-algorithms.

Floridi, Luciano, Josh COWls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 689-707.

Google LLC. 2018. *AI at Google: our principles*. June 07. Accessed December 21, 2018. https://www.blog.google/technology/ai/ai-principles/.

Habib, Kareem. 2017. PE 16-007 *Automatic vehicle control systems*. Accident Investigation, National Highway Traffic Safety Administration.

Hagendorff, Thilo. 2019. "The Ethics of Ethics: An Evaluation of Guidelines." *CoRR* 2.

Marshall, Aarian, and Alex Davies. 2018. "Uber's Self-Driving Car Saw the Woman It Killed, Report Says." *Wired*. 05 24. Accessed 04 05, 2019. https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/.

McNamara, Andrew, Justin Smith, and Emerson Murphy-Hill. 2018. "Does ACM's code of ethics change ethical decision making in software development?" *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering*, 729-733. Lake Buena Vista, Florida: Association for Computing Machinery.

Microsoft Corporation. n.d. Microsoft Design. Accessed January 16, 2019. https://www.microsoft.com/design/inclusive/.

Monteiro, Mike. 2017. "A Designer's Code of Ethics". Mule Design. Accessed January 16, 2018. https://muledesign.com/2017/07/a-designers-code-of-ethics.

O'Neill, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

Rawls, John. 1999. *A Theory of Justice*. 2. Belknap Press: An Imprint of Harvard University Press.

Sanger, David E, and Nicole Perloth. 2016. "A New Era of Internet Attacks Powered by Everyday Devices." *The New York Times*, October 16: accessed December 20, 2018, https://www.nytimes.com/2016/10/23/us/politics/a-new-era-of-internet-attacks-powered-by-everyday-devices.html.

Wakabayashi, Daisuke. 2018. "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam." *The New York Times*, March 19: accessed December 20, 2018, https://www.nytimes.com/2018/03/19/technology/uber-driver-less-fatality.html.

Williams, Andrew B. 2018. "The Potential Social Impact of the Artificial Intelligence Divide." *AAAI 2018 Spring Symposium on AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents* . Stanford, CA: Association for the Advancement of Artificial Intelligence.

<div></div>
R. Dumond, K. Dent i M. Kuniavsky
<div></div>
<i>Marc per aplicar sistemàticament l'ètica humanista quan s'utilitza la IA com a material de disseny</i>
<div></div>
Traducció al Català
<div></div>
PARAULES CLAU
<div></div>
Ètica, Tecnologia, Societat, Disseny amb IA.
<div></div>
RESUM
<div></div>

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

Els sistemes d'aprenentatge automàtic i d'IA cada vegada adquireixen més capacitats i s'utilitzen en més àmbits. I nosaltres, com a dissenyadors, creadors i investigadors, hem de tenir en compte les implicacions del seu ús, tant les positives com les negatives. En vista del que acabem d'exposar, els investigadors del Palo Alto Research Center (PARC) reconeixen que no es pot abaixar la guàrdia davant el dany que la intel·ligència artificial pot provocar en forma de discriminació intencionada o inadvertida, tracte injust o lesions que es poden infligir a persones o grups de persones. Com que els processos de decisió autònoms i emparats per IA podrien comportar efectes personals, socials i mediambientals negatius generalitzats, ens proposem adoptar una posició proactiva de defensa dels drets humans, de respecte de la intimitat dels individus, de protecció de la informació personal i de foment de la llibertat d'expressió i de la igualtat.

La tecnologia no és neutral per naturalesa i reflecteix les decisions i les concessions dels dissenyadors, els investigadors i els enginyers que la creen i la utilitzen en la seva feina. Els conjunts de dades sovint reflecteixen distorsions històriques. Les tecnologies d'IA que contracten persones, n'avaluen el rendiment laboral, proporcionen atenció sanitària i imposen moltes són exemples evidents de possibles àmbits en què es poden produir errors sistemàtics en els algoritmes que donin lloc a un tracte injust. Gairebé tota la tecnologia inclou concessions i encarna els valors i els judicis de les persones que l'han creat. Per això és imprescindible que els investigadors siguin conscients dels judicis de valor que fan i que siguin clars pel que fa al cas amb totes les parts implicades.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

INTRODUCCIÓ

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

1. La IA planteja molts desafiaments
Qualsevol tecnologia, des del foc fins a l'energia nuclear o les xarxes socials, comporta conseqüències inesperades, i com més poderosa sigui la tecnologia, més greus seran les conseqüències. És impossible impedir completament les conseqüències inesperades d'una nova tecnologia, però creiem que es poden crear pràctiques que, si més no en part, prevegin i atenuïn les que tenen un impacte negatiu directe i a curt termini sobre les persones que hi interactuen.

La intel·ligència artificial (en totes les seves formes, des de tasques senzilles com ara el reconeixement de patrons fins a altres de més com-

plexes com ara la planificació, la predicció i el raonament) és increïblement potent, de manera que les conseqüències inesperades que comporten els seus efectes són especialment preocupants. La nostra experiència com a dissenyadors d'algoritmes d'IA, d'experiències d'usuari amb la IA i de models impulsats per IA ens ha permès conèixer de primera mà els de-safiaments que comporta treballar amb aquest nou material de disseny i ser conscients de la importància de crear pràctiques ètiques de disseny que adverteixin els dissenyadors dels usos perillosos de la IA i alhora els donin eines per atenuar alguns dels seus possibles usos.

Només l'enumeració dels desafiaments que planteja la IA és descoratjadora per si mateixa, perquè qualsevol cosa que toquen els ordinadors (que és pràcticament tot) es converteix en un punt de possibles conseqüències negatives quan entra en joc la IA. Alguns dels desafiaments que planteja, com ara el dilema del tramvia per a la conducció autònoma o la tecnologia de reconeixement facial entrenada amb conjunts de dades distorsionats, són evidents. Altres, com ara l'impacte de l'automatització en els llocs de feina, l'explotació algorítmica de les distorsions cognitives i l'ús coercitiu de governs i empreses opressores, comencen a sorgir. La majoria dels desafiaments que planteja la IA són totalment imprevisibles, i segurament només els coneixerem quan s'hagin implementat completament les tecnologies que utilitzen.

Amb aquest estudi no busquem donar solucions als desafiaments actuals que planteja la IA, sinó més aviat oferir als dissenyadors implicats en el desenvolupament d'experiències impulsades per la IA un marc ètic que podran utilitzar per valorar les implicacions potencialment negatives de la seva feina. D'aquesta manera podran determinar per si mateixos perills específics de la seva pròpia feina i també podran notificar-los a l'organització per a la qual treballen. Relatem l'esforç que hem dedicat a la nostra feina al Palo Alto Research Center (anteriorment Xerox PARC). El PARC té una dilatada trajectòria a Silicon Valley i ha estat la punta d'espasa de moltes innovacions. Cada vegada veiem més aplicacions en el món real de la nostra feina i de la feina d'altres amb IA, i per això volem assegurar-nos que comprenem plenament l'impacte que podrien arribar a tenir les seves conseqüències negatives i estudiar amb deteniment maneres de pal·liar-les.

Una de les nostres principals prioritats és que quan hi ha determinats riscos, o quan s'assumeixen determinats supòsits, aquests s'han de fer palesos i s'han de comunicar a totes les parts implicades, incloses les afectades per la IA. El progrés de la tecnologia és una negociació continuada entre els seus creadors, els organitzadors que financen la feina i els usuaris finals, que són els consumidors -o els objectius- d'aquesta tecnologia. Quan es crea una tecnologia no s'ha de prendre simplement una única decisió ètica (a banda d'una decisió explícita de fer o no fer alguna cosa intencionadament per fer mal a les persones, perjudicar el medi ambient o minar l'ordre social; en aquest cas no hi ha procés i ens hauríem de resistir a plantejar-nos treballar en un projecte d'aquesta mena). Més aviat cal prendre decisions sobre el tema, l'objectiu, la transparència i l'experiència d'una tecnologia que es van presentant al llarg del projecte i que requereixen una consideració ètica, deliberació, debat i consens.

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

2. Pensem que amb els nous protocols i les llistes de verificació no n'hi ha prou

Els principis d'IA de Google, per exemple, són escanyolits (Google LLC 2018). Extenses disquisicions filosòfiques sobre el dilema del tramvia, en totes les seves possibles formes, només abracen una petita porció d'un tipus d'impacte negatiu d'una tecnologia determinada. Parlar-ne és com parlar de l'impacte negatiu d'inhalar vapors de benzina sense debatre sobre la resta d'efectes negatius de la benzina, des de la seva toxicitat fins a l'impacte mediambiental dels motors de combustió, passant per la seva inflamabilitat. Sí, la gent no ha d'inhal·lar benzina, però si ens centrem en això, perdem de vista la visió de conjunt i molts dels seus aspectes, que són realment importants.

Per norma general, amb una senzilla llista de verificació no n'hi ha prou. Tampoc és una qüestió de normes, sinó de processos: el procés de revisió del disseny, el procés d'estudi dels usuaris i el procés de recopilació de dades. Els processos han de preveure la supervisió de grups de diferents àmbits, així com la transparència de la recopilació de dades i dels algoritmes. El disseny ha d'incloure la capacitat de qüestionar els sistemes amb mecanismes clars per controlar-los. Els processos han d'incloure senyals d'avertència a tall de fusibles dels procediments que indiquin quan s'està entrant en àrees perilloses i que especifiquin què passa si un procés es descontrola.

K. DENT / R. DUMOND / M. KUNIAVSKY

3. Governament del procés de disseny

Creiem que, tal com passa amb la investigació mèdica, amb els processos de plantes químiques, amb la conducció en autopistes i amb moltes altres activitats diàries, també hi ha d'haver un model de governament per al disseny amb IA. Els projectes han de ser capaços d'entendre i preparar participants com ara científics de dades, científics socials, els responsables de fixar normes -dins i fora d'una organització- i dissenyadors de tota mena. El propòsit no és frenar la creació amb més burocràcia, sinó obligar a reflexionar sobre el tipus de coses que es poden trencar quan vas molt ràpid, quin pla hi ha per identificar quan s'estan trencant abans que es trenquin completament i com arreglar-les abans que es trenquin definitivament. El nostre procés beu d'altres plantejaments similars com el de l'Institutional Review Boards (IRBs), el del disseny participatiu, el de l'ètica del disseny i el de la Computer Professionals for Social Responsibility (CPSR).

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

UN CAS DE DISSENY INCLUSIU I INVESTIGACIÓ D'EXPERIÈNCIA D'USUARI

R. Dumond, K. Dent i M. Kuniavsky, amb els seus fills, a la casa de la família a la ciutat de Ann Arbor, Michigan, 1979.

Actualment hi ha un encès debat sobre les implicacions de la IA tant en els mitjans de comunicació com en l'àmbit acadèmic. Diverses conferències i tallers sobre temes relacionats amb la IA tracten la qüestió del bé. En aquestes trobades els diferents punts de vista tendeixen a considerar l'ètica d'una o dues maneres: com un desafiament tècnic (el problema de dotar de moralitat els agents intel·ligents per prendre decisions) o com una manera de frenar una poderosa força social negativa. Es tracta de debats realment encomiables, però a banda del problema (molt real) de la distorsió en les dades (O'Neill 2016), la realitat és que s'està parlant ben poc sobre els danys que els sistemes intel·ligents estan provocant i provocaran en el futur immediat i sobre la resposta pràctica que s'hi hauria de donar actualment. Per exemple, al simposi AAAI 2018 Spring Symposium on AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents, catorze del setze treballs presentats inclöïen estratègies per dotar de moralitat els agents en l'adopció de decisions i només dos, (Williams 2018; Dent 2018), tractaven l'impacte ètic dels sistemes d'IA en la vida de les persones. Això no ha de sorprendre si tenim en compte que es tracta d'una conferència tècnica, però gairebé tots els altres treballs preveïen un futur en què agents d'IA altament sofisticats participen plenament en l'àmbit de la vida humana. Actualment el sector està creant tecnologies d'IA molt menys sofisticades que tenen efectes nocius per a les persones i la societat; tanmateix, la majoria de creadors tenen altres prioritats. No només hem negligit la nostra responsabilitat d'imaginar totes les conseqüències que podria tenir la IA en la nostra vida, sinó que hem provocat danys en no ser crítics amb els riscos que els nostres sistemes d'IA poden comportar.

Una altra activitat relacionada amb la IA i l'ètica és la creació de directrius. Recentment el món acadèmic i l'empresarial han publicat diverses directrius. Novament, tal com destaca Hagendorff (2019): "...s'han escrit molt poques coses o cap sobre la implementació tangible de valors i objectius ètics". Hagendorff analitza quinze grups de directrius. Fa palès que no hi ha proves del seu impacte tangible, i destaca que un dels problemes és que aquestes inclouen termes com ara “justícia”, “transparència” i “centrat en la persona”, però no es defineix o es puntualitza què volen dir. A més, les directrius voluntàries tenen el problema inherent, pel seu caràcter voluntari, que no se'n pot obligar el compliment.

Un estudi publicat l'any passat (McNamara, Smith i Murphy-Hill 2018) es plantejava si les directrius ètiques publicades per l'Association for Computing Machinery tenen alguna influència en decisions ètiques de l'àmbit del software. Es plantejava als participants una sèrie de situacions i tot seguit se'ls preguntava quines decisions prendrien en determinades circumstàncies. Els investigadors no van detectar diferències estadística-ment significatives entre les respostes del grup d'estudi que havia rebut indicacions de tenir en compte les directrius i les d'un grup de control que no les coneixia.

Tot i amb això, creiem que publicar directrius ajuda a donar a conèixer els problemes existents i aporta arguments al debat. És útil exposar que disposar d'un bloc de principis referits a la IA en què la majoria estigués d'acord seria positiu per a la humanitat. Un dels millors exemples de directrius, AI4People (Floridi et al. 2018), propugna actituds que segurament generarien usos beneficiosos de la IA. Malgrat això, no esperem veure canvis concrets en les decisions i els comportaments de dissenyadors i investigadors com a conseqüència de l'existència de directrius.

TEMES DE DISSENY #35

ESTUDI DE CAS

188

189

Algunes directrius pràctiques no referides a la IA que s'utilitzen en el sector són realment reveladores. Microsoft ha creat un kit d'eines amb indicacions perquè els dissenyadors utilitzin la inclusivitat en els seus productes. És curiós que, tot i que Microsoft centra els seus esforços en l'accessibilitat, el kit d'eines nasqués de la IA. Quan el gegant informàtic treballava en Cortana, l'assistent d'IA de Microsoft, els dissenyadors van haver de fer front a les limitacions que impliquen les interfícies d'usuari basades en el llenguatge oral.

- Kit d'eines de disseny inclusiu de Microsoft*

Microsoft descriu el disseny inclusiu com “una metodologia de disseny que habilita i utilitza tot l'espectre de diversitat humana. I encara més important, això vol dir incloure-hi persones que tenen diferents punts de vista i aprendre d'elles.” (Microsoft Corporation s.d.)

Microsoft impulsa el disseny inclusiu centrant-se en tres elements fonamentals:

- Admetre l'exclusió*: El disseny inclusiu ens recorda que les nostres capacitats i experiències personals introdueixen una distorsió. Quan no tenim en compte altres experiències, i ens basem exclusivament en nosaltres mateixos per trobar solucions de disseny, exclom altres que tenen capacitats i experiències personals diferents.
- Aprendre de la diversitat*: Quan una situació i la capacitat de la persona no coincideixen, la gent s'adapta. Aquestes adaptacions permeten conèixer maneres de millorar els serveis.
- Solució per a un aplicada a molts*: El disseny per a discapacitats també beneficia altres usuaris que tenen una limitació circumstancial. El kit d'eines és un exemple de dispositiu dissenyat per a una persona amb un sol braç, però que també pot ajudar una persona amb una lesió de canell temporal o un pare que sosté un nadó.

Microsoft Inclusive demostra que el disseny inclusiu no s'hauria de considerar una penalització posterior al procés de disseny, sinó que més aviat hauria de formar part del procés de disseny. De forma anàloga, creiem que l'ètica forma part del procés de disseny amb IA, és a dir, que l'ètica en IA és un material de disseny.

2. Investigació d'experiència d'usuari

Un equip diversificat implica no tenir només en compte la raça i el sexe, sinó també la diversitat de gènere, la capacitat o discapacitat, la situació socioeconòmica, el nivell d'educació, la religió, el país d'origen, l'edat o el domini de la tecnologia, entre altres. Els equips amb una gran diversitat tenen fàcil accés a un ampli ventall d'experiències. Un grup en què tots els membres tinguin experiències vitals similars no obtindrà mai els mateixos resultats que un altre que es pugui basar en diferents punts de vista. Els equips diversificats, en què hi ha espai per a totes les veus, són capaços d'incorporar diferents maneres de veure el món amb un mínim esforç addicional.

L'exigua diversitat dels equips tècnics de Silicon Valley, sens cap dubte, ha contribuït al dany que la IA ha provocat en els últims anys. Però fins i tot tenint un equip diversificat, aconseguir empatitzar simplement a partir de l'experiència del teu propi equip té un límit. És inqüestionable que hi ha d’haver més diversitat en els equips tècnics, però més enllà d'això es poden plantejar arguments a favor d'una investigació d'experiència d'usuari més extensa.

Quan els dissenyadors pensen en investigació d'experiència d'usuaris, sovint pensen en els nostres clients. Dissenyadors i enginyers necessiten un canvi radical per incloure l'experiència de totes les parts implicades en el procés de disseny. El disseny d'experiència d'usuari d'un projecte d'IA no s'hauria de centrar només en la usabilitat d'una interfície ni a determinar si la interfície d'usuari és vàlida per a qualsevol que hi interactui (per exemple, si un botó “Enviar” hauria d'estar alineat a la dreta o a l'esquerra de la pantalla). El disseny d'IA ha d'incloure obligatòriament les experiències de les persones per a les quals aquesta pren decisions. Per exemple, ha de tenir en compte l'experiència de la persona a la qual es nega la llibertat condicional basant-se en models d'aprenentatge automàtic.

Mike Monteiro, dissenyador de Mule Design, va publicar el seu codi deontològic per a dissenyadors el 2017, un llibret vermell titulat *Design Ethics*. Monteiro, entre altres coses, sosté que no hi ha casos extrems en disseny. Segons la seva opinió, afirmar que determinats grups representen un cas extrem és marginar-los intrínsecament (Monteiro 2017). Monteiro utilitza Facebook com a exemple. La xarxa social afirma que té més de 2.000 milions d'usuaris, de manera que un 1% (una xifra que es podria considerar un cas extrem) equival a 20 milions de persones. Una decisió que ignori un 1% del total exclou i margina 20 milions de persones. Tal

com ens recorda Monteiro: “No són casos extrems. Són éssers humans i mereixen tot el nostre esforç”. (Monteiro 2017)

Cal tenir en compte la gravetat de les conseqüències. Que Netflix et recomani una pel·lícula horrible perquè has vist d'amagat *Bake off* et pot molestar, però te n'oblides després de navegar una estona per la plataforma de televisió. Tanmateix, ser atropellat per un vehicle autònom perquè no t'ha identificat correctament té conseqüències mortals (Marshall i Davies 2018). Quan parlem d'automatització hi ha moltes coses en joc.

CREACIÓ D'UN MARC I UN PROCÉS ÈTICS PER AL DISSENY AMB IA

Quan els dissenyadors i investigadors d'IA raonem sobre la nostra feina i sobre l'impacte que aquesta té en el món, hi ha alguns mètodes per centrar l'atenció i per analitzar amb esperit crític les possible conseqüències que són millors que uns altres. Els millors mètodes segurament generaran resultats verços, creïbles i més beneficiosos. La diferència entre una consideració superficial i un procés rigorós pot tenir conseqüències considerables. Tenint això en compte, els investigadors de PARC han creat una metodologia de disseny per treballar amb el material potencialment perillós d'IA.

Els codis ètics d'obligada aplicació, tal com proposem, tenen precedents. Un que és força similar al nostre plantejament i que el va inspirar és l'informe Belmont Report (Belmont 1979). L'informe, que va ser creat pel Govern dels Estats Units, analitza i presenta principis i directrius ètics per a la protecció dels subjectes humans d'estudis d'investigació. Una sèrie d'incidències que van ser de domini públic van dur a la creació de l'informe, el Tuskegee Syphillis Study, que va dur a terme el servei sanitari públic dels Estats Units entre el 1932 i el 1972. Ara l'estudi és cèlebre per les seves violacions de normes ètiques, ja que els investigadors van retirar deliberadament el tractament a pacients malalts.

Tràgicament i irònica, els investigadors del Tuskegee, i altres com ells, molt probablement es van convèncer a sí mateixos que estaven actuant correctament. Per a aquells que no ho veiessin tan clar, una breu i desencertada estimació moral amb unes conclusions que van presentar uns clars beneficis per a la societat sens dubte van aconseguir calmar les preocupacions. Actualment, els dissenyadors i investigadors amb IA estan en una situació similar. Els mitjans de comunicació lloen les enormes bondats de la tecnologia basada en IA. Menys visibles són les conseqüències negatives que s'han trobat i que han generat molt poc debat sobre el potencial d'altres danys. En l'actual cultura de disseny amb IA, fins i tot els professionals més benintencionats es mostren incapaçs d'entendre la responsabilitat que tenen quan creen noves aplicacions d'IA, fins i tot quan aquesta implica prendre decisions sobre vides humanes o interactuar amb humans de formes que els posin en perill (Wakabayashi 2018).

Al llarg de la història molts altres documents han fixat normes per regular l'actuació i aconseguir minimitzar el dany en les interaccions humanes, especialment quan es produeixen relacions desiguals de poder (govern davant individus, persones afortunades davant persones desafavorides, persones amb privilegis davant persones sense representació, persones formades davant persones sense qualificació, etc). Alguns d'aquests ara tenen el suport de la legislació internacional. Documents com el Codi de Nuremberg, la Declaració Universal dels Drets Humans i la Declaració de Hèlsinki van ser intents de restablir un ordre moral després d'episodis d'abusos dels drets humans. En tots aquests documents hem trobat alguns principis d'alt nivell que són aplicables a la nostra feina i que han guiat el nostre pensament:

- Respecte per les persones*
- Seguretat de la persona*
- Respecte del dret de la persona a la privacitat, la protecció de les dades, la llibertat d'expressió i la participació en la vida cultural.*
- Igualtat i imparcialitat*
- Els beneficis per a la societat no s'han d'aconseguir a costa d'altres*

DIRECTRIUS DEL DISSENY AMB IA

Atesa la nostra situació al cor de Silicon Valley als Estats Units, és inevitable que tinguem una perspectiva determinada sobre els problemes que la tecnologia planteja a la societat. Estem envoltats de les últimes innovacions

i tenim la possibilitat d'observar els efectes -positius i negatius- que tenen en diferents grups i subgrups de persones. Quan ens plantegem els potencials efectes d'una tecnologia nova no ens limitem a consultar una llista de normes ètiques esperant que el corrent ens dugui en la direcció correcta. Al món real l'ètica no és mai senzilla. L'excel·lència ètica comença fent front a les qüestions dures que duen a debats complexos que tracten sacrificis difícils. D'altra banda, tot el que es qüestiona i es debat ha de col·locar les consideracions ètiques en el context de l'ús previst d'una tecnologia determinada. Proposem un comitè d'experts de l'organització que ajudi els investigadors i dissenyadors a negociar els paisatges ètics dels seus projectes. Els membres del comitè segueixen la seva pròpia formació ètica i milloren les seves competències amb formació, lectures i debats centrats en temes d'actualitat, i també amb debats oberts sobre actes i incidents de tot el món que són rellevants des del punt de vista de l'ètica.

Tot seguit hi ha una sèrie de preguntes d'orientació amb què un comitè pot obrir debats sobre un projecte per dur la conversa cap a la complexitat, no per allunyar-se'n.

1. Llista de verificació provisional

Respectar el temps i l'atenció dels dissenyadors implica que les revisions ètiques no siguin gaire tedioses i que no se sol·licitin quan no sigui necessari. Una senzilla llista de verificació ajuda a valorar el dany potencial i el grau d'afectació.

Quan la IA aprèn de les dades existents, pregunta't:

Les dades del teu projecte representen persones o grups de persones?	Sí/No
Les dades contenen atributs personals (atributs especialment protegits com ara la raça, el sexe, la identitat de gènere, la capacitat o discapacitat, la posició, la situació socioeconòmica, el nivell d'educació, la religió o el país d'origen)?	Sí/No
L'objectiu és fer prediccions sobre el comportament de la gent? L'objectiu és classificar les persones o fer prediccions sobre aquestes d'alguna altra manera?	Sí/No
L'objectiu és prendre decisions sobre persones o grups de persones que podrien tenir repercussions destacables en la seva vida (per exemple, en el rendiment a la feina, en sentències judicials, en activitats fraudulentes, etc.)?	Sí/No
Quan la IA participa en espais físics on hi ha càmeres, micròfons, sensors o qualsevol altre element que enregistri l'activitat o l'aparença humana:	
Està en un lloc públic?	Sí/No
L'element que enregistra està ocult de qui pot ser enregistrat? Dit d'una altra manera, les persones podrien ser enregistrades sense saber-ho?	Sí/No
Es podria enregistrar la informació de membres de grups vulnerables (qualsevol subconjunt de persones desafavorides d'un grup de població, com ara nens, presos, refugiats o persones que pateixen algun tipus de discriminació)?	Sí/No
Hi ha maquinària pesada o maquinària que funciona a alta velocitat? Si la resposta a alguna de les preguntes anteriors és afirmativa	Sí/No
molt probablement caldrà més vigilància.	

2. Directrius generals

Quan un projecte requereix una dedicació més atenta, proposem un conjunt de preguntes per obrir un debat i una exploració dels danys potencials. No totes són pertinents per a tots els projectes. Les preguntes estan pensades per fer sorgir idees que els membres del comitè puguin utilitzar per formular preguntes sobre el projecte en qüestió. L'objectiu dels debats que es generen així és aconseguir que els dissenyadors es plantegin molts aspectes de la tecnologia. Organitzem les preguntes en diferents categories: dades humanes, impacte social i mediambiental, interacció física, actuació malintencionada i impacte ecològic. Algunes de les preguntes s'han adaptat a partir del compendi de principis FAT/ML (imparcialitat, rendició de comptes i transparència en l'aprenentatge automàtic) de l'organització (FAT/ML s.f.).

2.1. Dades humanes

És fonamental el respecte de les persones com a individus. La compilació de dades de caràcter personal i conductual no pot sobrepassar el que s'ha estipulat en el consentiment informat dels subjectes. Pensi com

podria afectar l'anàlisi de les dades a cadascun dels subjectes implicats. Si es tracta de dades de persones vulnerables (per exemple, una persona gran) o de persones sense plena autonomia (per exemple, criatures), s'han de protegir amb una atenció especial. Cal respectar la individualitat i l'autonomia de cada persona, i cal prendre mesures per protegir-la de qualsevol dany.
› *Ha adoptat les mesures adequades per protegir la informació tenint en compte que es tracta de dades sensibles?* Qualsevol informació personal d'identificació és informació potencialment sensible i, en molts casos, tal com passa amb la informació mèdica, necessita diversos nivells de manipulació quan és tractada.

› *Si el projecte és per a una entitat de caràcter públic, pot revelar les fonts de les seves dades?* D'on provenen les dades, qui les va compilar i amb quin objectiu ho va fer no només pot afectar la seva qualitat, sinó que també pot modificar la qualitat d'altres aspectes, incloses possibles distorsions.

› *Ha confirmat que les seves dades reflecteixin fidedignament la situació de la vida real del problema que intenta resoldre? Podria recórrer a fonts de dades alternatives?* Obtenir les dades adequades pot ser car i de vegades fins i tot impossible. Treballar amb unes dades perquè són les que hi ha disponibles és una pràctica sorprenentment habitual. Això ha derivat en un tracte injust de persones a causa de suposicions equivocades i conclusions errònies.

› *Ha comprovat la coherència interna de les dades (amb mostres aleatòries, per exemple)?* Ha d'informar d'algun problema a les parts implicades? Com comprovarà si hi ha distorsions explícites o implícites a les dades? Tingui en compte també subgrups de població. Faci les comprovacions adequades perquè no hi hagi distorsions estadístiques i pensi detingudament si les dades encaixen o no amb el fenomen de la vida real. És important comunicar totes les limitacions i suposicions a les parts implicades.

› *Els beneficis del seu disseny s'estendran a tota la població o hi podria haver subgrups que en quedessin inadvertidament exclosos?* En casos en què hi ha un risc conegut de discriminació, inclogui detalls de l'aplicació i les seves funcions al costat de mostres de les dades i els detalls de les seves fonts. Hi ha un procés per prendre mesures efectives que limitin la discriminació?

› *En projectes amb un considerable risc de provocar abusos dels drets humans, es poden enviar les dades a auditors externs independents?* La participació d'actors externs pot no ser convenient, ja que molts de nosaltres treballem per a empreses privades; tanmateix, quan és possible, la revisió independent del sistema i dels seus resultats és una bona verificació de la feina realitzada, especialment si el sistema pren decisions que podrien tenir un impacte significatiu en la vida de les persones.

› *Les parts implicades poden auditar el sistema?* Com a mínim, no hauria de crear sistemes de caixa negra que les persones afectades no puguin inspeccionar i revisar.

› *Hi ha un pla per comunicar les decisions sobre sacrificis, premisses dels projectes, mancances, índexs d'error, etc. a les parts implicades?* Durant el procés de disseny i desenvolupament, s'han de prendre decisions sobre premisses i sacrificis. Si no es documenten i comuniquen, no estaran mai en coneixement fora de l'equip de desenvolupament.

› *Hi ha un procés per corregir dades o rebatre decisions equivocades?* Els sistemes poden fallar i no tenen mai un 100% d'exactitud. Les persones que estiguin afectades pels sistemes han de tenir una manera de qüestionar els resultats.

› *Quan escau, hi ha una manera d'eliminar les dades si és necessari (per exemple, el dret a l'oblit tal com està contemplat a la UE i a l'Argentina?)*

2.2. Impacte social

La IA és una part cada vegada més destacada de molts aspectes de la nostra vida, inclosos certs processos de presa de decisions que podrien alterar-ne el curs. Els bons dissenyadors no consideren el possible impacte social com una ocurrència d'últim moment o com un impediment per al projecte. Més aviat pensen activament com impedir danys mentre es crea el sistema.

2.2.1. *Decisions d'IA que podrien distorsionar o alterar convencions socials*
› *El sistema beneficia més o menys alguns subgrups?* Cal fer una especial atenció als efectes dels errors en el sistema. Hi ha grups propensos a veure's perjudicats de manera més significativa que uns altres?
› *Cal pensar en els serveis, els productes i les indústries -i potser en els llocs de treball- que la implementació de l'algorítme/sistema podria substituir.*

K. DENT / R. DUMOND / M. KUNIAVSKY

191

TEMES DE DISSENY #35

ESTUDI DE CAS

190

proyecto de ese estilo). Más bien hay que tomar varias decisiones sobre el tema, el objetivo, la transparencia y la experiencia de una tecnología que se van presentando a lo largo del proyecto y que requieren una consideración ética, deliberación, debate y consenso.

2. Creemos que los nuevos protocolos y listas de verificación son insficientes

Los principios de IA de Google, por poner un ejemplo, son escualidos (Google LLC 2018). Extensas disquisiciones filosóficas sobre el dilema del tranvía, en todas sus posibles formas, no abordan más que una pequeña porción de un tipo de impacto negativo de una tecnología determinada. Hablar de ello es como hablar del impacto negativo de inhalar vapores de gasolina sin debatir sobre todos los otros efectos negativos que tiene la gasolina, desde su toxicidad, pasando por su inflamabilidad, hasta el impacto medioambiental de los motores de combustión. Sí, la gente no debe inhalar vapores de gasolina, pero si nos centramos en eso, perdemos de vista la visión de conjunto y muchos aspectos de la misma que son realmente importantes.

Por norma general, una simple lista de verificación no es suficiente. Tampoco es una cuestión de normas sino de procesos: el proceso de revisión del diseño, el proceso de estudio de los usuarios y el proceso de recopilación de datos. Los procesos deben prever la supervisión de grupos de distintos ámbitos así como la transparencia de la recopilación de datos y de los algoritmos. El diseño debe incluir la capacidad de cuestionar los sistemas con mecanismos claros para controlarlos. Los procesos deben incluir señales de advertencia a modo de fusibles de los procedimientos que señalen cuándo se está entrando en áreas peligrosas y que especifiquen qué sucede si un proceso se descontrola.

3. Gobernanza del proceso de diseño

Creemos que, al igual que sucede con la investigación médica, con los procesos de plantas químicas, con la conducción en autopistas y con muchas otras actividades diarias, también debe haber un modelo de gobernanza para el diseño con IA. Los proyectos deben ser capaces de comprender y preparar a participantes como, por ejemplo, científicos de datos, científicos sociales, los responsables de fijar normas –dentro y fuera de una organización– y diseñadores de todo tipo. El propósito no es frenar la creación con más burocracia sino obligar a reflexionar sobre el tipo de cosas que se pueden romper cuando vas muy rápido, qué plan hay para identificar cuándo se están rompiendo antes de que se rompan por completo y cómo arreglarlas antes de que se rompan definitivamente. Nuestro proceso bebe de otros planteamientos similares como el del Institutional Review Boards (IRBs), el del diseño participativo, el de la ética del diseño y el de la Computer Professionals for Social Responsibility (CPSR).

<p>UN CASO DE DISEÑO INCLUSIVO E INVESTIGACIÓN DE EXPERIENCIA DE USUARIO</p>	
<p>Actualmente existe un vivo debate sobre las implicaciones de la IA tanto en los medios de comunicación como en ámbito académico. Varias conferencias y talleres sobre temas relacionados con la IA abordan la cuestión del bien común. En dichos encuentros los distintos puntos de vista tienden a considerar la ética de una o dos maneras: como un desafío técnico (el problema de dotar de moralidad a los agentes inteligentes para tomar decisiones) o como una manera de frenar una poderosa fuerza social negativa. Se trata de debates realmente encomiables, pero aparte del problema (muy real) de la distorsión en los datos (O'Neill 2016), la realidad es que se está hablando muy poco sobre los daños que los sistemas inteligentes están provocando y provocarán en el futuro inmediato y sobre qué respuesta práctica habría que dar actualmente. Por ejemplo, en el simposio AAAI 2018 Spring Symposium on AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents, catorce de los dieciséis trabajos presentados incluían estrategias para surtir de moralidad a los agentes en la adopción de decisiones y solo dos, (Williams 2018; Dent 2018), abordaban el impacto ético actual de los sistemas de IA en la vida de las personas. Eso no debe sorprender si tenemos en cuenta que se trata de una conferencia técnica, pero casi todos los trabajos restantes preveían un futuro en el que agentes de IA altamente sofisticados participan plenamente en el ámbito de la vida humana. Actualmente el sector está creando tecnologías de IA mucho menos sofisticadas que tienen efectos nocivos para las personas y la sociedad; sin embargo, las prioridades de la mayoría de creadores son otras. No solo hemos negligido</p>	

TEMES DE DISSENY #35

nuestra responsabilidad de imaginar todas las consecuencias que podría tener la IA en nuestra vida, sino que hemos provocado daños al no ser críticos con los riesgos que nuestros sistemas de IA pueden comportar.

Otra actividad relacionada con la IA y la ética es la creación de directrices. Recientemente el mundo académico y el empresarial han publicado varias directrices. Nuevamente, como destaca Hagendorff (2019): “... se ha escrito muy poco o nada sobre la implementación tangible de valores y objetivos éticos”. Hagendorff analiza quince grupos de directrices. Pone de manifiesto que no hay pruebas del impacto tangible de las directrices, y destaca que uno de los problemas es que estas incluyen términos como “justicia”, “transparencia” y “centrado en la persona”, pero no se define o se puntualiza qué significan. Además, las directrices voluntarias tienen el problema inherente, por su carácter voluntario, de que no se puede obligar su cumplimiento.

Un estudio publicado el año pasado (McNamara, Smith y Murphy-Hill 2018) se planteaba si las directrices éticas publicadas por la Association for Computing Machinery tienen alguna influencia en decisiones éticas del ámbito del software. Se planteaba a los participantes una serie de situaciones y a continuación se les preguntaba qué decisiones tomarían en determinadas circunstancias. Los investigadores no detectaron diferencias estadísticamente significativas entre las respuestas del grupo de estudio que había recibido indicaciones para tener en cuenta las directrices y las de un grupo de control que no conocía dichas directrices.

Así y todo, creemos que publicar directrices ayuda dar a conocer los problemas existentes y aporta argumentos al debate. Es útil exponer que disponer de un bloque de principios referidos a la IA en el que la mayoría estuviera de acuerdo sería positivo para la humanidad. Uno de los mejores ejemplos de directrices, AI4People (Floridi et al. 2018), propugna actitudes que seguramente generarían usos beneficiosos de la IA. No obstante, no esperamos ver cambios concretos en las decisiones y los comportamientos de diseñadores e investigadores como consecuencia de la existencia de directrices.

Algunas directrices prácticas no referidas a la IA que se utilizan en el sector son realmente reveladoras. Microsoft ha creado un kit de herramientas con indicaciones para que los diseñadores utilicen la inclusividad en sus productos. Es curioso que, aunque Microsoft centra sus esfuerzos en la accesibilidad, el kit de herramientas nació de la IA. Cuando el gigante informático trabajaba en Cortana, el asistente de IA de Microsoft, los diseñadores tuvieron que hacer frente a las limitaciones que implican las interfaces de usuario basadas en el lenguaje oral.

1. Kit de herramientas de diseño inclusivo de Microsoft

Microsoft describe el diseño inclusivo como “una metodología de diseño que habilita y hace uso de todo el espectro de diversidad humana. Y aún más importante, eso significa incluir a personas que tienen distintos puntos de vista y aprender de ellas.” (Microsoft Corporation s.f.)

Microsoft impulsa el diseño inclusivo centrándose en tres elementos fundamentales:

1. Admitir la exclusión: El disseny inclusiu ens recorda que les nostres capacitats i experiències personals introdueixen una distorsió. Quan no tenim en compte altres experiències, i ens basem exclusivament en nosaltres mateixos per trobar solucions de disseny, excluem altres que tenen capacitats i experiències personals diferents.

2. Aprender de la diversidad: Cuando no coinciden una situación con la capacidad de la persona la gente se adapta. Estas adaptaciones permiten conocer maneras de mejorar los servicios.

3. Solución para uno aplicada a muchos: El diseño para discapacidades también beneficia a otros usuarios que tienen una limitación circunstancial. El kit de herramientas da un ejemplo de dispositivo diseñado para una persona con un solo brazo pero que también puede ayudar a una persona con una lesión de muñeca temporal o a un padre que sostiene a un bebé.

Microsoft Inclusive demuestra que el diseño inclusivo no debería considerarse una penalización posterior al proceso de diseño sino que más bien debería formar parte del proceso de diseño. Análogamente, creemos que la ética forma parte del proceso de diseño con IA, es decir que la ética en IA es un material de diseño.

2. Investigación de experiencia de usuario

Un equipo diversificado implica no tener solo en cuenta la raza y el sexo, sino también la diversidad de género, la capacidad o discapacidad, la situación socioeconómica, el nivel de educación, la religión, el país de origen, la edad o el dominio de la tecnología, entre otros. Los equipos con una gran

ESTUDIO DE CASO

194

diversidad tienen fácil acceso a un amplio abanico de experiencias. Un grupo en el que todos los miembros tengan experiencias vitales similares nunca obtendrá los mismos resultados que otro que pueda basarse en distintos puntos de vista. Los equipos diversificados, donde hay espacio para todas las voces, son capaces de incorporar distintas maneras de ver el mundo con un mínimo esfuerzo adicional.

La exigua diversidad de los equipos técnicos de Silicon Valley ha contribuido, sin duda alguna, al daño que la IA ha provocado en los últimos años. Pero aun teniendo un equipo diversificado, lograr empatizar simplemente a partir de la experiencia de tu propio equipo tiene un límite. Es incuestionable que debe haber más diversidad en los equipos técnicos, pero más allá de eso pueden plantearse argumentos a favor de una investigación de experiencia de usuario más extensa.

Cuando los diseñadores pensamos en investigación de experiencia de usuarios, a menudo pensamos en nuestros clientes. Diseñadores e ingenieros necesitan un cambio radical para incluir la experiencia de todas las partes implicadas en el proceso de diseño. El diseño de experiencia de usuario de un proyecto de IA no debería centrarse solo en la usabilidad de una interfaz ni en determinar si la interfaz de usuario es válida para cualquiera que interactúe con ella (por ejemplo, si un botón de “Enviar” debería estar alineado a la derecha o a la izquierda de la pantalla). El diseño de IA debe incluir obligatoriamente las experiencias de las personas para las que esta toma decisiones. Por ejemplo, debe tener en cuenta la experiencia de la persona a la que se le niega la libertad condicional en base a modelos de aprendizaje automático.

Mike Monteiro, diseñador de Mule Design, publicó su código deontológico para diseñadores en 2017, un librito rojo titulado *Design Ethics*. Monteiro sostiene, entre otras cosas, que no existen casos extremos en diseño. En su opinión, afirmar que determinados grupos representan un caso extremo es marginarlos intrínsecamente (Monteiro 2017). Monteiro utiliza Facebook como ejemplo. La red social afirma tener más de 2000 millones de usuarios, por lo que un 1% (una cifra que podría considerarse un caso extremo) equivale a 20 millones de personas. Una decisión que ignore a un 1% del total excluye y margina a 20 millones de personas. Tal como nos recuerda Monteiro: “No son casos extremos. Son seres humanos y merecen todo nuestro esfuerzo”. (Monteiro 2017)

Hay que tener en cuenta la gravedad de las consecuencias. Que Netflix te recomiende una película horrible porque has mirado a escondidas *Bake off* puede molestar, pero lo olvidas después de navegar un rato por la plataforma de televisión. Sin embargo, ser atropellado por un vehículo autónomo porque no te ha identificado correctamente tiene consecuencias mortales (Marshall y Davies 2018). Cuando hablamos de automatización hay mucho en juego.

<p>CREACIÓN DE UN MARCO Y UN PROCESO ÉTICOS PARA EL DISEÑO CON IA</p>	
<p>Cuando los diseñadores e investigadores de IA razonamos sobre nuestro trabajo y sobre el impacto que este tiene en el mundo, hay determinados métodos de centrar la atención y de analizar con espíritu crítico las posibles consecuencias que son mejores que otros. Los métodos mejores seguramente generarán resultados veraces, creíbles y más beneficiosos. La diferencia entre una consideración superficial y un proceso riguroso puede tener consecuencias considerables. Teniendo eso en cuenta, los investigadores del PARC han creado una metodología de diseño para trabajar con el material potencialmente peligroso de IA.</p> <p>Los códigos éticos de obligada aplicación, como proponemos, no carecen de precedentes. Uno que guarda una estrecha semejanza con nuestro planteamiento y que lo inspiró es el informe Belmont Report (Belmont 1979). El informe, que fue creado por el Gobierno de EEUU, analiza y presenta principios y directrices éticos para la protección de los sujetos humanos de estudios de investigación. Una serie de incidentes que fueron de dominio público llevaron a la creación del informe, el Tuskegee Syphilis Study, que llevó a cabo entre 1932 y 1972 el servicio sanitario público de EEUU. Ahora el estudio es célebre por sus violaciones de normas éticas, pues los investigadores retiraron deliberadamente el tratamiento a pacientes enfermos.</p> <p>Trágica e irónicamente, los investigadores del Tuskegee y otros como ellos muy probablemente se convencieron a sí mismos de que estaban actuando correctamente. Para los que no lo tuvieran tan claro, una breve y desacertada estimación moral con unas conclusiones que presentaron unos</p>	

K. DENT / R. DUMOND / M. KUNIAVSKY

claros beneficios para la sociedad sin duda lograron calmar las preocupaciones. Actualmente, los diseñadores e investigadores con IA se encuentran en una situación similar. Los medios de comunicación alaban las enormes bondades de la tecnología basada en IA. Menos visibles son las consecuencias negativas que ya se han visto y que han generado muy poco debate sobre el potencial de otros daños. En la actual cultura de diseño con IA, hasta los profesionales más bienintencionados se muestran incapaces de comprender la responsabilidad que tienen al crear nuevas aplicaciones de IA, incluso cuando esta implica tomar decisiones sobre vidas humanas o interactuar con humanos de maneras que ponen en peligro su vida (Wakabayashi 2018).

A lo largo de la historia muchos otros documentos han fijado normas para regular la actuación de manera que se minimice el daño en las interacciones humanas, especialmente cuando se dan relaciones desiguales de poder (gobierno ante individuos, personas afortunadas ante personas desfavorecidas, personas con privilegios ante personas sin representación, personas formadas ante personas sin cualificación, etc). Algunos de ellos ahora están respaldados por la legislación internacional. Documentos como el Código de Núremberg, la Declaración Universal de los Derechos Humanos y la Declaración de Helsinki fueron intentos de restablecer un orden moral tras episodios de atropellos de los derechos humanos. En todos esos documentos hemos encontrado algunos principios de alto nivel que son aplicables a nuestro trabajo y que han guiado nuestro pensamiento:

1. Respeto por las personas

2. Seguridad de la persona

3. Respeto del derecho de la persona a la privacidad, la protección de los datos, la libertad de expresión y la participación en la vida cultural

4. Igualdad e imparcialidad

5. Los beneficios para la sociedad no deben ser a costa de otros

<p>DIRECTRICES DEL DISEÑO CON IA</p>	
<p></p>	

Dada nuestra situación en el corazón de Sillicon Valley de Estados Unidos, es inevitable que tengamos una perspectiva determinada sobre los problemas que plantea la tecnología a la sociedad. Estamos rodeados por las últimas innovaciones y tenemos la posibilidad de observar los efectos –positivos y negativos– que tienen en distintos grupos y subgrupos de personas. Cuando nos planteamos los potenciales efectos de una tecnología nueva no nos limitamos a consultar una lista de normas éticas esperando que la corriente nos lleve en la dirección correcta. En el mundo real la ética casi nunca es sencilla. La excelencia ética empieza por afrontar cuestiones duras que llevan a debates complejos que abordan sacrificios difíciles. Por otra parte, todo cuanto se cuestiona y se debate debe colocar las consideraciones éticas en el contexto del uso previsto de una tecnología determinada. Proponemos un comité de expertos de la organización que ayude a los investigadores y diseñadores a negociar los paisajes éticos de sus proyectos. Los miembros del comité siguen su propia formación ética y mejoran sus competencias con formación, lecturas y debates centrados en temas de actualidad, así como con debates abiertos sobre actos e incidentes de todo el mundo que son relevantes desde el punto de vista de la ética.

A continuación figura una serie de preguntas de orientación con las que un comité puede abrir debates acerca de un proyecto para llevar la conversación hacia la complejidad, no para alejarse de ella.

<p><i>1. Lista de verificación provisional</i></p> <p>Respetar el tiempo y la atención de los diseñadores implica que las revisiones éticas no sean demasiado tediosas y que no se soliciten cuando no sea necesario. Una simple lista de verificación ayuda a valorar el daño potencial y el grado de afectación.</p> <p>Cuando la IA aprende de los datos existentes, pregúntate:</p> <p>¿Los datos de tu proyecto representan a personas o a grupos de personas?</p> <p>¿Los datos contienen atributos personales (atributos especialmente protegidos como la raza, el sexo, la identidad de género, la capacidad o discapacidad, la posición, la situación socioeconómica, el nivel de educación, la religión o el país de origen)?</p> <p>¿El objetivo es hacer predicciones sobre el comportamiento de la gente?</p> <p>¿El objetivo es clasificar a las personas o hacer predicciones sobre ellas de algún otro modo?</p>	<p>Sí/No</p> <p>Sí/No</p> <p>Sí/No</p> <p>Sí/No</p> <p>Sí/No</p> <p>Sí/No</p>
--	---

195

¿El objetivo es tomar decisiones sobre personas o grupos de per-
sonas que podrían tener repercusiones destacables en
su vida (por ejemplo, en el rendimiento en el trabajo, en
sentencias judiciales, en actividades fraudulentas, etc.)?
Cuando la IA participa en espacios físicos donde hay cámaras, mi-
crófonos, sensores o cualquier otro elemento que registre la actividad o la
apariencia humana:
¿Está en un lugar público?
¿El elemento que registra está oculto a ojos de quien puede ser
registrado? Dicho de otro modo, ¿podrían las personas
ser grabadas sin saberlo?
¿Se podría registrar la información de miembros de grupos
vulnerables (cualquier subconjunto de personas des-
favorecidas de un grupo de población, como niños,
presos, refugiados o personas que sufren algún tipo
de discriminación)?
¿Hay maquinaria pesada o maquinaria que funciona a alta ve-
locidad?
Si la respuesta a alguna de las anteriores preguntas es afirmativa
muy probablemente hará falta mayor vigilancia.

2. Directrices generales

Cuando un proyecto requiere una dedicación más atenta, propon-
emos un conjunto de preguntas para abrir un debate y una exploración de
los daños potenciales. No todas son pertinentes para todos los proyectos.
Las preguntas están pensadas para hacer surgir ideas que los miembros
del comité puedan utilizar para formular preguntas sobre el proyecto en
cuestión. Los debates que así se generen tienen por objetivo lograr que los
diseñadores se planteen muchos aspectos de la tecnología. Organizamos
las preguntas en distintas categorías: datos humanos, impacto social y
medioambiental, interacción física, actuación malintencionada e impacto
ecológico. Algunas de las preguntas se han adaptado a partir del compendio
de principios FAT/ML (imparcialidad, rendición de cuentas y transparencia
en el aprendizaje automático) de la organización (FAT/ML s.f.).

2.1. Datos humanos

Es fundamental el respeto de las personas como individuos. La re-
copilación de datos de carácter personal y conductual no puede sobrepasar
lo estipulado en el consentimiento informado de los sujetos. Piense en cómo
podría afectar el análisis de los datos a cada uno de los sujetos implicados.
Si se trata de datos de personas vulnerables (por ejemplo, ancianos) o de
personas sin plena autonomía (por ejemplo, niños), deben protegerse con
especial atención. Debe respetarse la individualidad y la autonomía de cada
persona, y deben tomarse medidas para protegerla de cualquier daño.
› ¿Ha adoptado las medidas adecuadas para proteger la información teniendo
en cuenta que se trata de datos sensibles? Toda información personal
de identificación es información potencialmente sensible y, en mu-
chos casos, como sucede con la información médica, necesita varios
niveles de manipulación cuando es tratada.

› Si el proyecto es para una entidad de carácter público, ¿puede revelar las fuentes
de sus datos? De dónde provienen los datos, quién los recopiló y con
qué propósito lo hizo puede afectar no solo a la calidad de los datos
sino que también puede modificar la calidad de otros aspectos,
incluyendo posibles distorsiones.

› ¿Ha confirmado que sus datos reflejen de manera fidedigna la situación de la
vida real del problema que intenta resolver? ¿Podría recurrir a fuentes
de datos alternativas? Obtener los datos adecuados puede ser caro y
en ocasiones incluso imposible. Trabajar con unos datos porque son
los datos que hay disponibles es una práctica sorprendentemente
habitual. Ello ha derivado en un trato injusto de personas a causa
de suposiciones equivocadas y conclusiones erróneas.

› ¿Ha comprobado la coherencia interna de los datos (con muestras aleatorias,
por ejemplo)? ¿Debe informar de algún problema a las partes implica-
das? ¿Cómo comprobará si hay distorsiones explícitas o implícitas
en los datos? Tenga en cuenta también a subgrupos de población.
Realice las comprobaciones adecuadas para que no haya distorsiones
estadísticas y piense detenidamente si los datos encajan o no con
el fenómeno de la vida real. Es importante comunicar todas las
limitaciones y suposiciones a las partes implicadas.

› ¿Los beneficios de su diseño se extenderán a toda la población o podría haber
subgrupos que quedaran inadvertidamente excluidos? En casos en los
que existe un riesgo conocido de discriminación, incluya detalles

de la aplicación y sus funciones junto con muestras de los datos y
detalles de las fuentes de dichos datos. ¿Existe un proceso para
tomar medidas efectivas que limiten la discriminación?

› En proyectos con un considerable riesgo de provocar abusos de los derechos hu-
manos, ¿se pueden enviar los datos a auditores externos independientes?
La participación de actores externos puede no ser conveniente,
pues muchos de nosotros trabajamos para empresas privadas; sin
embargo, cuando es posible, la revisión independiente del sistema
y de sus resultados es una buena verificación del trabajo realizado,
especialmente si el sistema toma decisiones que podrían tener un
impacto significativo en la vida de las personas.

› ¿Las partes implicadas pueden auditar el sistema? Como mínimo, no debería
crear sistemas de caja negra que las personas afectadas no puedan
inspeccionar y revisar.

› ¿Existe un plan para comunicar las decisiones sobre sacrificios, premisas de
los proyectos, carencias, índices de error, etc. a las partes implicadas?
Durante el proceso de diseño y desarrollo, ha tenido que tomar
decisiones sobre premisas y sacrificios. Si no se documentan y comu-
nican, nunca estarán en conocimiento fuera del equipo de desarrollo.

› ¿Existe un proceso para corregir datos o rebatir decisiones equivocadas? Los
sistemas pueden fallar y nunca tienen un 100% de exactitud. Las
personas que se vean afectadas por los sistemas deben tener modo
de cuestionar los resultados.

› Cuando procede, ¿existe un modo de eliminar los datos cuando sea necesario
(por ejemplo, el derecho al olvido tal como está contemplado en la UE y
en Argentina)?

2.2. Impacto social

La IA es una parte cada vez más destacada de muchos aspectos de
nuestra vida, incluidos ciertos procesos de toma de decisiones que podrían
alterar su curso. Los buenos diseñadores consideran el posible impacto
social no como una ocurrencia del último momento o como un impedimento
para un proyecto. Más bien piensan activamente en cómo impedir daños
mientras se crea el sistema.

2.2.1. Decisiones de IA que podrían distorsionar o alterar convenciones sociales

› ¿El sistema beneficia más o menos a algunos subgrupos? Hay que prestar una
atención especial a los efectos de los errores en el sistema. ¿Hay
grupos propensos a verse perjudicados de manera más significativa
que otros?

› Hay que pensar en los servicios, productos e industrias –y tal vez empleos– a
los que la implementación del algoritmo/sistema podría sustituir. ¿Qué
impacto tendrá en la sociedad? Aunque no pueda hacer nada, podría
seguir siendo indicado aplicar la innovación a largo plazo, pero la
perspectiva que tiene desde el interior le permite hacer aportaciones
o sugerencias para paliar los efectos negativos que el sistema tendrá
en las comunidades o en otros sistemas y grupos.

› ¿Cuando se utilizan datos humanos, los beneficios sobrepasan los riesgos de las
personas implicadas? Tenga en cuenta los beneficios de un proyecto y
el daño potencial que podrían sufrir los afectados por dicho proyecto.
Por ejemplo, investigadores de Stanford crearon una aplicación de
visión que predice las preferencias sexuales analizando imágenes de
caras. Se trata de una tecnología que prácticamente no ofrece ventaj-
as pero sí grandes costes potenciales en imparcialidad y equidad.
The Economist dijo al respecto: “En zonas del mundo donde ser gay
no está aceptado socialmente o es ilegal, dicho algoritmo podría ser
una grave amenaza para la seguridad” (Economist 2017).

› ¿Las decisiones que genera un sistema algorítmico son explicables a las per-
sonas afectadas por dichas decisiones? Las explicaciones ayudan en
la validación de resultados y aumentan la confianza en el sistema.

› Ejercicio útil: imagínese que usted mismo es el destinatario de los resultados del
sistema. ¿Son todos igualmente justos desde todos los puntos de vista?
(Rawls 1999). Cuando diseñe un sistema, póngase en la situación de
ser la persona afectada por las decisiones y predicciones del mismo.
No podemos eliminar todos los sesgos que nos son consustanciales,
pero imagínese a usted mismo con distintas capacidades naturales y
una posición distinta en la sociedad con identidades de género, raza
y nacionalidad diferentes.

2.3. Impacto medioambiental

El consumo de recursos de un sistema ya creado puede ser un as-
pecto ignorado del diseño.

› Tenga en cuenta el impacto energético de los recursos informáticos necesarios

para el proyecto. ¿Es posible minimizar dicho impacto? ¿El tratamiento
informático se puede realizar de otra forma para utilizar menos
energía? ¿Los recursos asignados al proyecto encajan con los re-
quisitos o son excesivos?

› ¿El algoritmo/sistema/diseño fomenta el comportamiento no sostenible? La
novedad de las cápsulas de café, por ejemplo, ha hecho que tomar
un café en cualquier momento sea extremadamente práctico, pero las
cápsulas no suelen ser reciclables y terminan en vertederos.

2.4. Interacción física

Los principales motivos de preocupación que despierta la IA están
relacionados con el respeto de los derechos y la privacidad de las personas.
El riesgo de lesiones físicas también puede ser motivo de preocupación. En
términos generales, los usuarios deberían conocer los posibles riesgos y los
creadores deben pensar detalladamente en las potenciales consecuencias
accidentales que puede tener interactuar con la tecnología.

› ¿Existe el equivalente a un consentimiento informado para el uso? El consen-
timiento informado implica que los usuarios comprenden totalmente
los hechos y las consecuencias que comporta su participación. Un
simple aviso legal y los densos acuerdos de uso son lo contrario de
un consentimiento informado.

› ¿Las personas que interactúen con la tecnología comprenderán sus capaci-
dades y sus límites? En mayo de 2016 los medios de comunicación
difundieron ampliamente la noticia del accidente de un coche Tesla
que funcionaba en modo de conducción automática (Habib 2017). Al
parecer el conductor tenía una fe excesiva en las capacidades del
coche: utilizó el modo de conducción autónoma en condiciones para
las que no estaba diseñado, lo que provocó su muerte.

› ¿El diseño minimiza los riesgos potenciales?

› ¿Las señales de aviso se corresponden con la gravedad de los peligros?

2.5. Uso indebido y actuación malintencionada

Una preocupación que suele pasarse por alto es el uso indebido de la
tecnología. El uso indebido puede ser consecuencia de un error, de un exceso
de confianza en la tecnología o de personas malintencionadas que intentan
sabotearla. En todos los casos se debe procurar paliar los posibles daños y
poner las restricciones adecuadas para limitar el uso indebido.

› ¿Existen medidas de protección para impedir un uso fuera del previsto (uso in-
debido intencionado o accidental del diseño)? En este caso el accidente
del Tesla vuelve a ser un buen ejemplo. Tesla habría podido diseñar
los controles de manera que los conductores no pudieran utilizarlos
allí donde no es seguro.

› ¿El sistema ha sido protegido adecuadamente para impedir la manipulación
exterior? Si se utilizan dispositivos físicos, ¿están debidamente pro-
tegidos del pirateo o el sabotaje? Por ejemplo, en 2016, dispositivos
de IoT, incluidos monitores para bebés, fueron pirateados y lanzaron
un ataque a gran escala contra una gran parte de internet (Sanger
y Perlroth 2016).

2.6. Tras la implementación

La responsabilidad de un diseño o un sistema no termina cuando
se entrega o se implementa el proyecto. Aunque deben realizarse todos los
esfuerzos posibles para evitar resultados inesperados, disponer de un plan
de extinción antes de entregar el proyecto al cliente puede evitar confusiones
sobre cómo resolver consecuencias imprevistas.

› ¿Las personas y los grupos pueden aplicar medidas efectivas de remedio
y reparación? Entre ellas puede haber, por ejemplo, crear procesos
claros de reparación de efectos o decisiones adversos.

› ¿Cómo serán los procesos de denuncia y de recurso?

› ¿Existe un plan que indique qué hacer si el proyecto presenta consecuencias
imprevistas? Este puede formar parte del plan de mantenimiento y
debería incluir planes de supervisión tras el lanzamiento.

CONCLUSIÓN

Con estos principios y directrices queremos incluir el debate sobre la ética
de la IA como un componente natural de los proyectos de IA diseñados y de-
sarrollados en el PARC. Queremos que nuestros procesos de tener en cuenta
la ética y de limitar los posibles riesgos sean habituales entre los equipos
que trabajan con IA. Creemos que hablando de las posibles consecuencias,

conociendo a fondo las experiencias de las partes implicadas y creando un
espacio para poner a revisión nuestros sesgos fomentamos una competencia
innata de ingeniería y diseño éticos.

	BIOGRAFÍAS
	<div><div><div><div><div></div><div>Kyle Dent</div></div><div><div></div><div>Palo Alto Research Center</div></div></div></div></div>
	<div><div><div><div><div></div><div>Richelle Dumond</div></div><div><div></div><div>Palo Alto Research Center</div></div></div></div></div>
	<div><div><div><div><div></div><div>Mike Kuniavsky</div></div><div><div></div><div>Palo Alto Research Center</div></div></div></div></div>

Mike Kuniavsky es creador, investigador y diseñador de experiencias de usuario. Con
veinte años de experiencia en desarrollo de productos digitales, Mike diseña productos,
procesos de negocio y servicios a la vanguardia del cambio tecnológico.
Antes de entrar en el PARC, fue uno de los fundadores de varias empresas dedicadas a las
experiencias de usuario, como ThingM, que diseña y fabrica productos omnipresentes de
informática y productos de internet de las cosas, y Adaptive Path, una conocida asesoría
de diseño. Está especializado en interacciones entre múltiples dispositivos, en diseño
de servicios en la nube y en diseño de productos de hardware conectados a servicios en
la nube. A lo largo de su carrera ha trabajado en análisis de redes sociales, electrónica
de consumo, dispositivos, recuperación de imágenes, LED RGB y servicios financieros.
Ha trabajado con algunas de las principales empresas tecnológicas del mundo, como
Samsung, Sony, Nokia, Whirlpool y Qualcomm, para las que ha diseñado nuevos productos,
ha coordinado estrategias de producto y ha creado una cultura de diseño y desarrollo
centrada en el usuario.
Es autor de *Observing the User Experience: A Practitioner’s Guide to User Research y
Smart Things: Ubiquitous Computing User Experience Design*, que se utilizan habitual-
mente como libros de texto en varias universidades. Obtuvo una doble licenciatura en
Informática y Estudios de Vídeo de la Universidad de Michigan.

FIGURAS

Fig. 1. En el diseño y la implementación de sistemas de IA, las decisiones que se toman
en relación al destinatario evidente tendrán repercusiones en otras partes. Todas las
partes afectadas son tan destinatarias del resultado de un diseño como el diseñador, el
cliente o lo que se ha venido en denominar tradicionalmente el «usuario». Es importante
que el diseño de sistemas de IA incluya las experiencias de todas las personas afectadas.
Ilustración: Ying Xiao.

Fig. 2. Los procesos deben incluir una supervisión por parte de grupos de distintos sectores
así como transparencia en la recopilación de datos y en los algoritmos. El diseño debe
incluir la capacidad de cuestionar los sistemas con mecanismos claros para controlarlos.
La transparencia debe mantenerse tras la implementación, controlando regularmente el
rendimiento y el funcionamiento real del sistema. Ilustración: Ying Xiao.

REFERENCIAS

Ver listado completo de referencias en la página 188.