AI & SOCIETY (2023) 38:2071–2079
https://doi.org/10.1007/s00146-020-01110-y

STUDENT FORUM

AI transparency: a matter of reconciling design with critique

Tomasz Hollanek[1]

Received: 28 March 2020 / Accepted: 29 October 2020 / Published online: 17 November 2020

Abstract
In the late 2010s, various international committees, expert groups, and national strategy boards have voiced the demand to 'open' the algorithmic black box, to audit, expound, and demystify artificial intelligence. The opening of the algorithmic black box, however, cannot be seen only as an engineering challenge. In this article, I argue that only the sort of transparency that arises from critique—a method of theoretical examination that, by revealing pre-existing power structures, aims to challenge them—can help us produce technological systems that are less deceptive and more just. I relate the question of AI transparency to the broader challenge of responsible making, contending that future action must aim to systematically reconcile design—as a way of concealing—with critique—as a manner of revealing.

Keywords Critical theory · Critical thinking · Transparency · Responsibility · Self-awareness · Design theory

1 Preliminaries

1.1 AI transparency

In the age of ubiquitous computing, we are surrounded by objects that incorporate artificial intelligence solutions. We interact with different kinds of AI without realizing it—using online banking systems, searching for YouTube clips, or consuming news through social media—not really knowing how and when AI systems operate. Corporate strategies of secrecy and user interfaces that hide traces of AI-driven personalization combine with the inherent opacity of deep learning algorithms (whose inner workings are not directly comprehensible to human interpreters) to create a marked lack of transparency associated with all aspects of emerging technologies. It is in response to the widespread application of AI-based solutions to various products and services in the late 2010s that multiple expert groups—both national and international—have voiced the demand to 'open' the algorithmic black box, to audit, expound, and demystify AI. They claim that to ensure that the use of AI is ethical, we must design emerging systems to be transparent, explainable, and auditable.[1]

The opening of the algorithmic black box, however, cannot be seen only as an engineering challenge. It is critique, as the underside of making, that prioritizes unboxing, debunking the illusion, seeing through—to reveal how an object *really* works. Critique—grounded in the tradition of Critical Theory and practiced by cultural studies, critical race theory, queer theory, as well as decolonial theory scholars, among others—moves beyond the technical detail to uncover the desires, ideologies, and social relations forged into objects, opening the black boxes of history, culture, and progress. In what follows, I argue that the calls for technological transparency demand that we combine the practice of design with critique. I relate the question of AI transparency to the broader challenge of responsible making, contending that future action must aim to systematically reconcile design—as a way of concealing—with critique—as a manner of revealing.

1.2 Levels of opacity

Many companies sell simple analytics tools as artificial intelligence, as something that supposedly supplants human intelligence to deliver better results. What is advertised as an

✉ Tomasz Hollanek
th536@cam.ac.uk

[1] Centre for Film and Screen, University of Cambridge, Cambridge, UK

[1] The High-Level Expert Group on AI convened by the European Commission presented its *Ethics Guidelines for Trustworthy Artificial Intelligence* in the early 2019. The document identifies several key characteristics of a system that can be deemed trustworthy, which include *transparency,* defined as a form of traceability of data, operations, and business models that shape the end product (AI HLEG

'AI solution' often relies on simple data analysis performed by human analysts. The use of metaphors and simplifications obfuscates human labor, labor that is outsourced, hidden away, in an invisible and immaterial factory, in a different part of the globe. According to Ian Bogost (2015), the 'metaphor of mechanical automation' is nothing more than a well-directed, but misleading, masquerade (n. pag.). Although the metaphor is only an approximation, a distortion, or even a 'caricature,' it convincingly plays the role of an accurate depiction of the whole. The term *artificial intelligence*, elusive, misleading and with a definition that has changed over time, forms the basis of the marketing stunt. Technology companies rely on abstract, overly schematic representations to simplify reality and arrive at an easily digestible, prepacked idea of the object, one that misrepresents the object's essence and overlooks its true composition, but also satiates the end user's curiosity.

Other systems, as a matter of deliberate practice, incorporate complex data processing and machine learning surreptitiously. Shoshana Zuboff (2019) observes that the influence these systems have on our decision-making is 'designed to be unknowable to us' (p.11); that company strategies of misdirection serve as 'the moat that surrounds the castle and secures the action within' (p.65.)—a way for corporations like Google or Facebook to protect their secrets and mislead the public.

These systems *could*, in theory, be designed to be more 'knowable.' Elements of the user interface could, for example, flag up when algorithmic operations are influencing the user's decision-making. Just like labels inform the consumer about the product's contents, the interfaces of Facebook and YouTube could announce to users that the information delivered by the platforms is algorithmically curated. Considering that only 24% of US college students realize Facebook automatically prioritizes certain posts and hides others (Powers 2017), such a feature would definitely be relevant. But the challenge of transparency at a time of unprecedented technological complexity cannot be approached only as a matter of failed (or indeed *successful*, depending on your position) communication.

The 'opacity' of machine learning algorithms refers, after all, not only to 'institutional self-protection and concealment,' but also, as Burrell (2016) points out, to 'the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation' (p.2). The fundamental lack of transparency of systems that incorporate AI solutions relates not only to convoluted storytelling devised by marketing teams, or misleading interfaces and user experience design, but also—and most importantly—an emerging form of making brought about by the automation of cognitive tasks themselves.

Considering this level of opacity of AI systems, a team of researchers from MIT's Intelligence and Decision Technologies Group developed a neural network named, suggestively, Transparency by Design Network (Mascharka et al. 2018) that not only performs 'human-like reasoning steps' to answer questions about the contents of images, but also 'visually renders its thought process' as it solves problems, allowing human analysts to interpret its decision-making. It is a deep learning system carrying out a sort of *unboxing* on itself, incorporating *explainability* in its very operations: a realization of the most common understanding of 'transparency' in the context of contemporary AI research.

1.3 Transparency and critique

The challenge of technological transparency that has now attracted the attention of policymakers has constituted a concern for cultural studies scholars, media theorists, and design philosophers for decades. In the early 1990s, the philosopher Paul Virilio (1994) noted that once 'synthetic vision' becomes a reality and the human subject is excluded from the process of observation of images 'created by the machine for the machine,' the power of statistical science to appear objective (and thus persuasive) will be 'considerably enhanced, along with its discrimination capacities' (p.75). Another prominent media theorist Friedrich Kittler (2014) expressed concerns about modern media technologies that are 'fundamentally arranged to undermine sensory perception.' Kittler wrote about a 'system of secrecy' based on user interfaces that 'conceal operations necessary for programming' and thus 'deprive users of the machine as a whole,' to suggest that perceiving the imitation, penetrating through the illusion of software 'that appears to be human,' is a fundamental challenge in the age of global-scale computing (pp. 221–24).

Transparency, as Adrian Weller (2017) has poignantly noted, is an ambiguous term that will mean different things to different people. For the user, a transparent system will 'provide a sense for what [it] is doing and why,' while for an expert or regulator, it will enable auditing 'a prediction or decision trail in detail' (p.3). The calls for technological transparency have thus been filtered down to reach various stakeholder groups with different meanings and representing different interests. But what seems consistent in all of

Footnote 1 (continued)

2019). Another interdisciplinary group of experts working under the auspices of IEEE (the world's largest professional association for electronic and electrical engineers) published, in 2019, the first edition of *Ethically Aligned Design*, a set of actionable recommendations on how to align design practices with society's values and principles, stressing that the 'standards of transparency, competence, accountability, and evidence of effectiveness should govern the development of autonomous and intelligent systems' (p. 5).

these diverse takes on the development, use, and regulation of technology is that transparency is framed as a matter of design. In what follows, I problematize this claim, arguing that design, in the most fundamental sense, relies on concealment and obfuscation. I contend that only the sort of transparency that arises from critique—a method of theoretical examination that, by revealing pre-existing power structures, aims to challenge them—can help us produce technological systems that are less deceptive and more just.

2 Design as blackboxing

2.1 Art and artifice

Coined in 1956 by John McCarthy, the term 'artificial intelligence' had its critics among those who attended the Dartmouth Conference (which famously established the field of AI); Arthur Samuel argued that 'the word artificial makes you think there's something kind of phony about this, […] or else it sounds like it's all artificial and there's nothing real about this work at all' (in: McCorduck 2004, p. 115). The historian of AI Pamela McCorduck notes that while other terms, such as 'complex information processing,' were also proposed, it was 'artificial intelligence' that endured the trial of time. According to her, it is 'a wonderfully appropriate name, connoting a link between art and science that as a field AI indeed represents' (p. 115). She is referring indirectly here to the origins of the word *artificial*; in Latin, *artificialis* means 'of or belonging to art,' while *artificium* is simply a work of art, but also a skill, theory, or system.

When the philosopher Vilém Flusser traced the etymology of the word 'design' in his *The Shape of Things: A Philosophy of Design* (1999), he referred to this relationship between art and artifice to argue that all human production, all culture, can be defined as a form of trickery. Flusser rejects the distinction between art and technology, and goes back to these ancient roots: the Greek for 'trap' is *mechos* (mechanics, machine); the Greek *techne* corresponds to the Latin *ars*; an *artifex* means a craftsman or artist, but also a schemer or trickster—to demonstrate that in their essence all forms of making are meant to help us 'elude our circumstances,' to cheat our own nature. Culture itself becomes a delusion brought about by means of design—a form of self-deception that makes us believe we can free ourselves from natural restrictions by producing a world of artifice. From doors to rockets, from tents to computer screens, from pencils to mechanized intelligences, Flusser selects his examples to show that, ultimately, any involvement with culture is based on deception: sometimes 'this machine, this design, this art, this technology is intended to cheat gravity, to fool the laws of nature' (ch.1, n. pag.)—and sometimes to trick ourselves into thinking we control both gravity and the laws of nature. In that sense, art and technology are representative of the same worldview in which cultural production must be deceptive/artful enough to enable humans to go beyond the limits of what is (humanly) possible.

Flusser refers to the act of weaving to explain the 'conspiratorial, even deceitful' (ch.18) character of design. In the process of carpet production, he points out, knotting is meant to deny its own warp, to hide the threads behind a pattern, so that anyone stepping on the finished rug perceives it as a uniform surface, according to the designer's plan. He offers weaving as one of the primordial forms of cultural production to embody trickery, but the same holds true for any form of design. The trick is always based on misdirection, shifting the end user's attention from the material to the application, from the current state of things to emerging possibilities and new futures. Designing is a methodical way of crafting alternative realities out of existing materials—a process of *casting* the intended shape onto the chosen fabric so as to create a new possibility. The material used in that process must, so to speak, dematerialize: it has to disappear from view and give way to the new object—to abstract the end result from the point of origin and the labor process. By obfuscating some components while exhibiting others, 'ideal' design enables an end user's cognitive efficiency.

2.2 Patterns, layers, and repetitions

For Flusser, any product of human making is both an object and an obstacle—Latin *objectum*, Greek *problema*—or, more specifically, any object is also an 'obstacle that is used for removal of obstacles' (ch.9). To move forward, we solve problems that lie ahead and obstruct our way; we produce objects that help us remove these obstacles; but the same objects turn into obstacles for those that come after us. In other words, since the results of human problem-solving are stored in objects, progress involves obfuscation and forgetting. We come up with a solution and, with time, this singular idea turns into a pattern; others use the already established template to produce new, more complex structures and these structures turn into new patterns, covering up previous layers of design with new design. To expedite the process of production, to advance, to move faster, the designer turns to these conventions and templates, choosing from a menu of preprogrammed options—or abstracting new rules based on previous patterns. And as the complexity of the production process increases, the reliance on patterns grows too. New design always depends on previous design, and this ultimate dependence on patterns and abstractions complicates understanding the process in its totality.

In the age of ubiquitous computing, speaking of *obfuscation by design* becomes of particular importance. In 2015, Benjamin Bratton called his model of the new kind

of layering brought about by planetary-scale computation 'the Stack':

> 'an accidental megastructure, one that we are building both deliberately and unwittingly and is in turn building us in its own image' (p.5).

New technologies 'align, layer by layer, into something like a vast, if also incomplete, pervasive if also irregular, software and hardware *Stack*' (p.5). This makes it hard to perceive the Stack's overarching structure, indeed, to see it *as* design, however incidental. Today, we produce new technologies, new objects, to see, know, and feel more, to register what is normally hidden from our view, meanwhile, creating complex systems based on multiple, invisible layers and algorithmic operations whose effects are not always comprehensible even to the designers themselves.

2.3 Automations and automatisms

In her comprehensive account of what she calls 'surveillance capitalism,' Shoshana Zuboff points out the dangers of technological illusion—'an enduring theme of social thought, as old as the Trojan horse' (p.16)—that serves the new economic project in rendering its influence invisible. Surveillance capitalism claims 'human experience as free raw material for translation into behavioral data,' and turns that data into 'prediction products' that can later be sold to advertisers (p.8). Echoing the work of philosophers such as Bernard Stiegler (2014, 2015) or Antoinette Rouvroy (2016), Zuboff argues that the ultimate goal of this new form of capitalism is 'to automate us,' by reprogramming our behavior and desires. Various internet platforms that dominate the market prompt us to action, influence our decision making, relying on big data analyses of our preference trends online. Automated systems create statistical models to profile users, tracing any emerging patterns in countless interactions with digital products; patterns turn into further abstractions, new models that are later reflected in new products and solutions, which end up 'automating' us, guiding our decision-making without our knowing.

But is this process specific to AI-enhanced personalization under surveillance capitalism? Bratton has recently argued that what 'at first glance looks autonomous (self-governing, set apart, able to decide on its own) is, upon closer inspection, always also decided in advance by remote ancestral agents and relays, and is thus automated as well' (2019, loc.345, n. pag.). Any decision taken now relies on multiple decisions taken in the past; new design depends on previous design; a new object coalesces from an aggregation of old solutions. Culture is an amalgamation of such objects—objects that, ironically, become obstacles because they are meant to enable our cognitive efficiency. A tool becomes an obstacle because the results of our problem-solving and labor are already stored within it; a tool must never be seen as a tool, as its use must be intuitive—it must remain imperceptible; any new tool meant to advance the process is made with existing tools, and so the emerging layering of design in the Anthropocene makes it harder to distinguish between tool and fabric. Extending this to the ongoing automation of cognitive tasks in the age of ubiquitous computing, the phenomenon takes on new scale.

This is why the emerging need for transparency refers not so much to company politics of disinformation or algorithmic black boxes, as to the very essence of our culture, as a process of knowledge production, pattern formation, and concealment. Particular problems caused by the widespread adoption of automated decision-making systems, such as algorithmic bias, can have specific, targeted, solutions in the form of new policy, engineering standards, or better education. But a shift of focus from the particular to the total is more than an exercise in theory—it makes us realize that transparency has never been at the heart of our making, that design has always been a form of blackboxing. There is, in that sense, something deeply anti-cultural about transparency. Or, putting it differently, there is nothing *natural* about transparency by design: we have been programmed to cover up as we make, not the opposite.

The ongoing transformation of lived experiences into data is a new analytical paradigm that demands our intervention, truly calls for an 'unboxing,' an excavation of processes and data trails. But the *opening* of the algorithmic black box cannot be viewed only as a technical issue—precisely because any solution is, first and foremost, a result of *cultural blackboxing*. While contemporary debates on AI focus on transparency as a direct response to the opacity of algorithms, what we are in need of are approaches that aim to 'unbox' new technologies as objects–obstacles, solutions that aim towards cognitive automation, products that store the results of problem-solving performed 'by remote ancestral agents,' and that can thus perpetuate injustices via automatically accepted patterns and norms.

3 Critique as unboxing

3.1 Apparent transparencies

Among entries on subjects such as theology, economics, and medicine, Denis Diderot and Jean le Rond d'Alembert included in their *Dictionary of the Sciences, Arts, and Crafts*, entries on artisanal practices that detail the individual steps in the processes of production adopted in clockmaking, tailoring, woodworking, and many others. One such entry focuses on the making of artificial flowers: the first plate (Fig. 1) depicts a dozen workers scattered across the

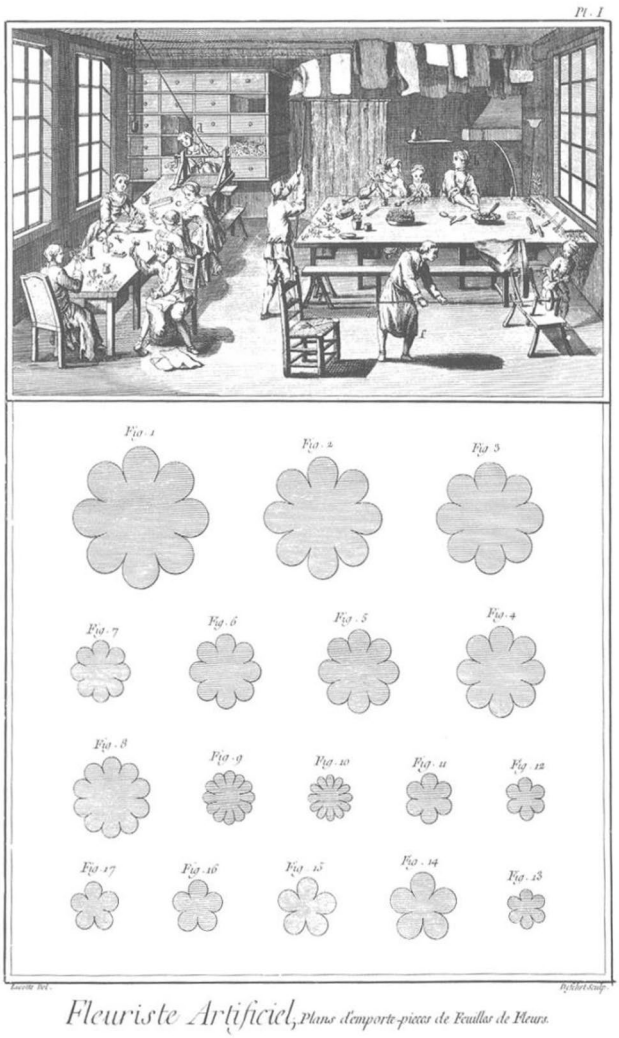

Fig. 1 Maker of artificial flowers, 1765 (https://hdl.handle.net/2027/spo.did2222.0001.451)

main workshop area, performing different tasks at various stages of manufacturing, while following pages of illustrations showcase the most popular templates used to emboss specific petal shapes onto fabric, with a final plate celebrating the finished commodity.

By bringing to view the backstages of production, the Encyclopedia was essentially *un*designing, reversing the process of 'conspiratorial weaving' described by Flusser. Now, in an age of growing technological complexity, shaped by significant degrees of cognitive automation, there is a need for a similar undesigning of new technologies. The artist Todd McLellan's photographs (Fig. 2) that document his multiple attempts at taking various objects apart are a suggestive illustration of this challenge in the age of extreme technological complexity. We might dissemble our smartphone, but learning what is hidden beneath the interactive surface of the touchscreen will never give us an indication

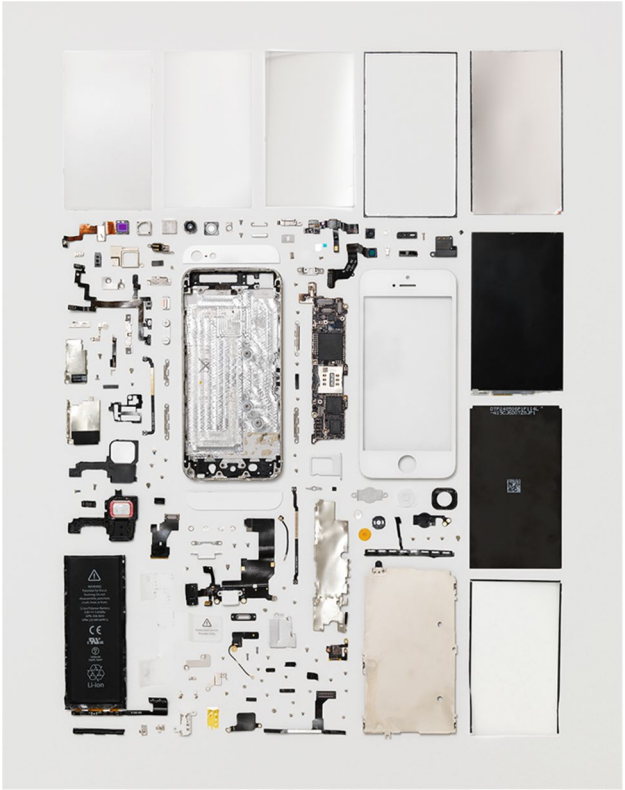

Fig. 2 Todd McLellan, Things Come Apart, 2013 (courtesy of the artist)

of how the device really works and, more significantly, in whose interest. The meaningless innards of the device become symbolic of the contestable quality that transparency really is—if we think of it as a condition for, or indeed a guarantee of, understanding. Critical undesigning cannot be confused with a simple act of reverse engineering. There can be transparency without critique, or *apparent* transparency: but a sort of transparency that does not arise from critical processes of unboxing is unlikely to advance comprehension.

In his lecture on black boxes, Galloway (2010) relates Marx's idea of descent into 'the hidden abode of production' (p.7), as a means of uncovering capital relations forged into commodities, to 'traditions of critical inquiry, in which objects were unveiled or denaturalized to reveal their inner workings—from Descartes's treatise on method […] to the Freudian plumbing of the ego' (p.5). Based on the assumption that the surface is merely a superficial facade to be penetrated by means of critique, these theories prioritized the interior and perceived objects as 'mystical black boxes waiting to be deciphered to reveal the rationality (of history, of totality) harbored within' (p.3).

For the purpose of this article, critique is understood as a broad set of methodologies, grounded in the tradition of Critical Theory, that perform a metaphorical dismantling of

objects to reveal how hidden and immaterial layers of design reflect social and economic structures—and how the power relations these structures generate become the sources of injustice, oppression, and exploitation. Critique's ultimate goal is to uncover and challenge the system(s) that objects of design engender; revelation is conceived of as the condition necessary for resistance—and systemic transformation. Looking beyond individual design flaws (and fixes), critique points to those 'ancestral relays' that automate our thinking—to patterns, repetitions, and automatisms so deeply ingrained, *weaved into* the fabric of our culture, that they become imperceptible—in particular to those who do not experience the injustices resulting from the adoption of already established patterns.

3.2 Critique in the age of AI

A former YouTube employee, Guillaume Chaslot, has coined the term *Algotransparency* to describe an experiment in which he investigates the terms appearing most frequently in the titles of videos recommended by YouTube. A program developed by Chaslot and his team traces thematic patterns in YouTube recommendations to prove there exists a systemic bias that promotes controversial clips. His research suggests Google's platform indeed 'systematically amplifies videos that are divisive, sensational and conspiratorial' (Lewis 2018)—that the recommendations are not related to the individual user's interests (as the company claims), but rather—exploit controversy to boost clickability. *Algotransparency* attempts to unbox the logic of YouTube's copyright-protected recommendation algorithm without directly looking into the system's black box, concentrating only on the *effects* of its activity. This specific experiment gives a good indication of where we should be directing our attention: focusing not so much on *how* YouTube operates, as on *why* it works at all. This question extends beyond the technicality of the algorithm, to more widely interrogate the forces orchestrating our consumption of digital goods and whose interests they serve—what Zuboff calls surveillance capitalism, or what Stiegler refers to as hyperindustrialism.

Ian Bogost (2015) has argued that the illusion of automation in technology—the trick that misdirects our attention from essential questions about human decision-making incorporated into emerging systems—breaks down 'once you bother to look at how even the simplest products are really produced.' In 2014 he collaborated with Alexis Madrigal to analyze Netflix's recommendation system and demonstrate that the platform's operations are distributed among so many different agents—including human curators who hand-tag all Netflix content—'that only a zealot would call the end result an algorithm'(Bogost 2015). Many experiments and critical projects try to achieve something similar: debunk the illusion of software by exposing AI as processual and collaborative, tracing the results of data analysis back to human decisions, biases, and labor.

In their *Anatomy of an AI System*, for instance, Crawford and Joler (2018) present a figurative dissection of Amazon's Echo device that brings to view the invisible mechanisms and dynamisms that the product encapsulates (Fig. 3). The detailed mapping of various objects and agents, as well as multiple layers of interaction between those elements, constitutes a representation of the system as composed not only of hardware and software, data and computation, but also human labor and planetary resources. Critique in the age of extreme technological complexity is as much about dissecting and penetrating, as it is about charting the invisible and immaterial terrains of interaction, analysis, consumption, and computation; mapping wider relations between energies, influences, and resources under surveillance capitalism—patterns of exploitation of both people and environments that the production of objects/obstacles entails.

In another project, Crawford teamed up with the artist Trevor Paglen to carry out what they call an *archeology of datasets,* such as ImageNet, used in machine learning to train AI to recognize specific elements of images—sets that can also become sources of bias inscribed into emerging systems. By e*xcavating* the datasets' underlying structures, Crawford and Paglen (2019) aim to reveal their implicit meanings:

> 'we have been digging through the material layers, cataloguing the principles and values by which something was constructed, and analyzing what normative patterns of life were assumed, supported, and reproduced.'

The blackboxed Earth, a world deeply transformed by ubiquitous computing, by various layers of what Bratton calls the Stack, demands this form of unearthing. Successful critique in the age of AI exposes the technology as relying on human cognition and decision-making; more broadly, critique reveals the constellation of objects/obstacles as products of layered problem-solving, a flawed process that is *necessarily* tainted by pre-existing patterns and abstractions, biases and beliefs. Prioritizing reflection over efficiency, critique becomes a methodical way of resisting our reliance on patterns—patterns that allow us to move faster, but that can also harbor previously made assumptions—about gender (Costanza-Chock 2018) or race (Benjamin 2019) as particularly poignant examples—and thus perpetuate, rather than challenge, pre-existing forms of injustice.

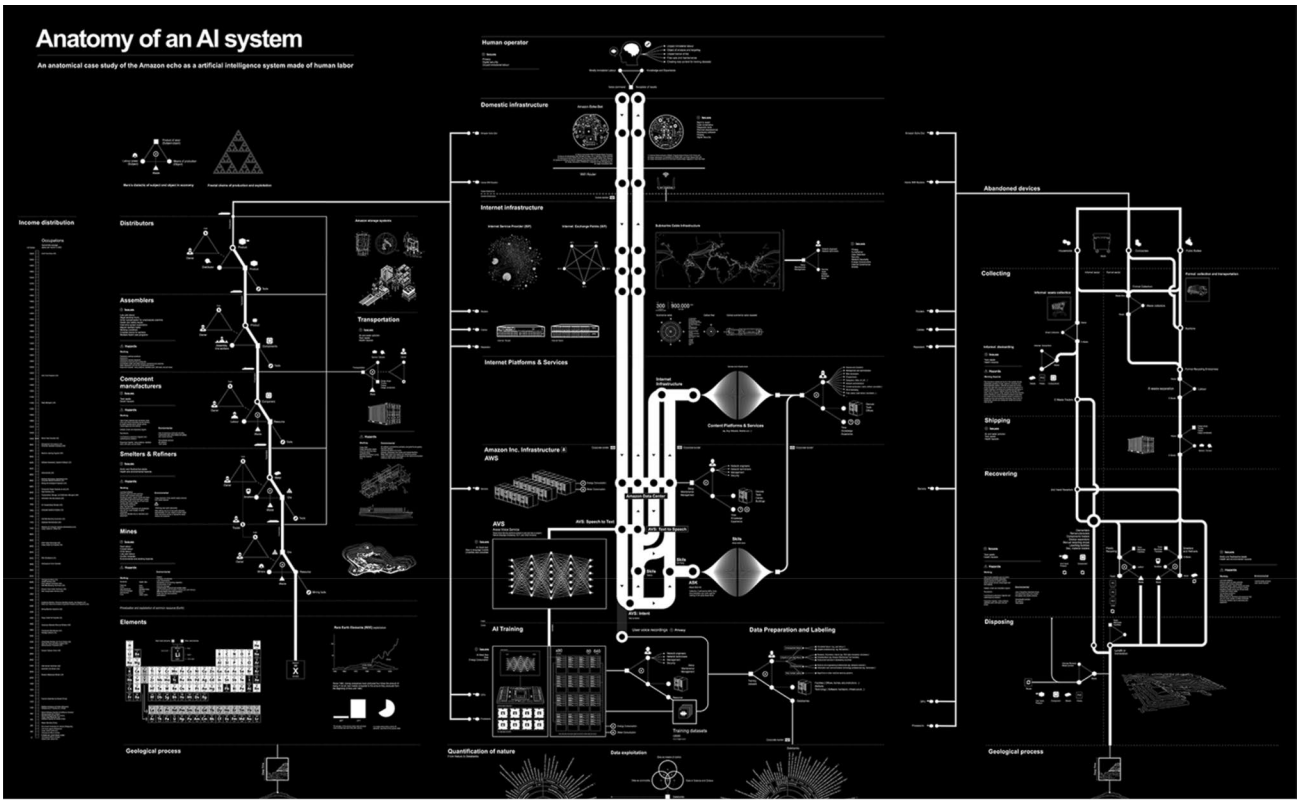

Fig. 3 Kate Crawford and Vladan Joler, Anatomy of an AI System, 2018 (courtesy of the artists)

4 Reconciling design with critique

In keeping up with the societal demand for transparent AI, the big players of the tech industry have been introducing changes in their engineering standards and organizational structures, hiring ethicists and policy specialists to cooperate with their product development teams. In 2016, Microsoft established the Aether Committee, a body of senior advisors to the company's leadership, that provides guidance 'on rising questions, challenges, and opportunities with the development and fielding of AI technologies' (2020), and oversees the work of other teams 'to ensure that AI products and services align with Microsoft's AI principles'—which include transparency and accountability. In 2017, DeepMind set up its Ethics and Society team 'to guide the responsible development and deployment of AI.' The team composed of 'ethicists and policy researchers' collaborates with the company's AI research team 'to understand how technical advances will impact society, and find ways to reduce risk.' Smaller industry players who decide to follow suit, but cannot afford to establish their own ethics 'departments,' begin to enlist the help of 'ethical auditing' companies; Glassbox, for example, is a tech consultancy startup, founded in 2018, that aims to 'provide clarity to the black box' by analyzing software products for signs of bias and training the client company's employees about systemic injustices. This way, elements of critique that reveal potential implications of human decision-making in design are supposed to become part of the production pipeline.

These sites of interaction between 'humanists' and 'technologists' in the industry—even if, in some cases, they amount to nothing more than backdrops for press releases—deserve our attention. Specifically, they require of us a comprehensive rethinking of what satisfies our desire for transparency in the age of extreme technological complexity. Can a system of checks and balances in the industry, an ongoing negotiation between blackboxing and unboxing, lead to anything more than design that *anticipates* critique? Is critique from within the industry necessarily a compromise and, therefore, nothing more than another step in the process of production of objects that are also obstacles? Must critique be external to the process of designing to remain genuine?

If design is, fundamentally, blackboxing and automation, and critique is unboxing that aims to reverse the process of 'conspiratorial weaving,' then we could conclude that these two sides of human activity are in stark opposition to one another—that design is incompatible with critique. Instituting a real change in the way we move forward must focus precisely on tackling this ostensible impossibility—on reconciling design with critique, progress with suspension,

production with reflection. The calls for *transparency by design* require that the 'making' itself be reinvented to incorporate critique.

If a systemic transformation depends on a new-found compatibility between design and critique, then the designers of emerging systems should turn to already existing, alternative design practices to learn what combining the processes of making and critique might entail. Anthony Dunne and Fiona Raby, who have been pioneers in the field of 'critical design' since the late 1990s, have been advocating for design understood as 'critical thought translated to materiality' (2013, p.35). An object of design in their framing must become a critical challenge—as much for the designer, as for the user: 'it encourages people to question, in an imaginative, troubling, and thoughtful way, everydayness and how things could be different.' (p.189) More recently, Ratto (2011) coined the term 'critical making' to describe a design process that focuses on 'the act of shared construction itself as an activity and a site for enhancing and extending conceptual understandings of critical sociotechnical issues' (p.254)—with those who take part as the agents of critique.

For Dunne and Raby, design has become 'so absorbed in industry, so familiar with dreams of industry, that it is almost impossible to dream its own dreams' (p.88). The suggestion is that the challenge lies not so much in the incompatibility between making and critique, as between the futures imagined by the industry and the dreams functioning outside of it. While elements of critique—gender critique and critical race theory in particular—seem to have already penetrated sections of the industry in the form of the mentioned ethics auditing services, reconciling the critique of surveillance capitalism and hyperindustrialism, the forces behind most of today's innovation, with the design of new technologies within corporate structures appears a considerable (and counterintuitive) feat. Perhaps this is where we should direct our attention: more research is needed on how practices such as critical design and critical making can influence the process of AI design; how critique can be operationalized within the industry to challenge industrial values and visons, including the idea of 'progress' itself.

In a recently published volume on the practice of 'undesign' (McNamara and Coombs 2019), Cameron Tonkinwise's essay proposes what he calls 'anti-progressive' design as a means of interrogating the designers' internalized desire for 'progress.' While users, as he rightly observes, are willing 'to unlearn and relearn modes of interaction' if what is new is also 'easier and more convenient, and hopefully more effective and pleasurable' (p.76), it is the designers who should learn how 'not to prefer progress, or how to prefer what does not feel like progress' (p.81). Tonkinwise argues it is the designers' duty

> 'to find a way to pursue the destructively preferable without casting the resulting change as progress: what is preferable are futures that no longer appear to be mere advancements of what currently exists' (p.81).

This is to say that the responsibility for future action lies, primarily, with the designers: the human makers who, as products of a specific culture, are being increasingly challenged to become aware of their own biases and automatisms that predetermine their actions and choices. For designers, combining design with critique, rather than an attempt at making *things* transparent, would constitute an attempt at becoming self-conscious.

Flusser argued that a renewed form of culture would have to be a culture 'aware of the fact that it was deceptive' (ch.1). To rethink technological transparency, we should first recognize that the designer has always been a trickster laying out traps, technologizing misdirection to pave the way forward. All aspects of design—practical, political, moral—would have to reflect this awareness: as we move forward, we must acknowledge that any problem solved now will also form a trap for those coming after us.

References

Algotransparency. https://algotransparency.org/index.html. **(Accessed 20 Jan 2020)**.

AI HLEG (2018) Requirements of Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Transparency. **(Accessed 10 Nov 2019)**

Benjamin R (2019) Race after technology: abolitionist tools for the new Jim code. Polity, Cambridge

Bogost I (2015) The cathedral of computation. In: The Atlantic. https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/. **(Accessed 14 Jun 2019)**

Bratton B (2015) The stack: on software and sovereignty. MIT Press, Cambridge

Bratton B (2019) The terraforming. Kindle ed.

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512

Costanza-Chock S (2018) Design justice A.I., and escape from the matrix of domination. J Design Sci. https://doi.org/10.21428/96c8d426 **(Accessed 20 Nov 2019)**

Crawford K, Joler V (2018) Anatomy of an AI System: The Amazon Echo as an anatomical map of human labor, data and planetary resources. https://anatomyof.ai. **(Accessed 20 Nov 2019)**

Crawford K, Paglen T (2019) Excavating AI: the politics of training sets for machine learning. https://excavating.ai. **(Accessed 20 Dec 2019)**

DeepMind. https://deepmind.com/safety-and-ethics. **(Accessed 02 Feb 2020)**

Ethically Aligned Design, First Edition (2019) https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf. **(Accessed 10 Dec 2019)**

Flusser V (1999) The shape of things: a philosophy of design. Apple Books ed.

Galloway A (2010) Black box, black bloc. A lecture given at the New School in New York City on April 12, 2010. https://cultureandcommunication.org/galloway/pdf/Galloway,%20Black%20Box%20Black%20Bloc,%20New%20School.pdf. **(Accessed 14 Jun 2019)**

Glassbox (2019) https://glassboxinc.com. **(Accessed 30 Oct 2019)**

Kittler F (2014) There is no software. In: The truth of the technological world: essays on the genealogy of presence. Stanford University Press, Stanford

Lewis P (2018) "Fiction is outperforming reality": how YouTube's algorithm distorts truth. https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth. **(Accessed 25 Feb 2018)**

Mascharka D, Tran P, Soklaski R, Majumdar A (2018) Transparency by design: closing the gap between performance and interpretability in visual reasoning. Proc IEEE Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2018.00519

McCorduck P (2004) Machines who think. A K Peters Ltd., Natick

Microsoft AI principles (2019). https://www.microsoft.com/en-us/ai/our-approach-to-ai. **(Accessed 1 Dec 2019)**

Microsoft (2020) Our approach to responsible AI. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4phnK, **(Accessed 02 Feb 2020)**

Powers E (2017) My news feed is filtered?: Awareness of news personalization among college students. Digit Journal 5(10):1315–1335

Ratto M (2011) Critical making: conceptual and material studies in technology and social life. Inf Soc 27(4):252–260. https://doi.org/10.1080/01972243.2011.583819 **(Accessed 1 Dec 2019)**

Rouvroy A (2016) Algorithmic governmentality: radicalization and immune strategy of capitalism and neoliberalism? In: La Deleuziana 3: Life and Number, 30–36

Stiegler B (2014, 2015) Symbolic misery, Volume 1 and Volume 2. Polity Press, Cambridge

Tonkinwise C (2019) "I prefer not to": anti-progressive designing. In: McNamara A, Coombs G et al (eds) Undesign. Critical practices at the intersection of art and design. Routledge, New York

Virilio P (1994) The vision machine. Indiana University Press and British Film Institute, Bloomington

Weller A (2017) Transparency: motivations and challenges. https://arxiv.org/abs/1708.01870. **(Accessed 02 Dec 2019)**

Zuboff S (2019) The age of surveillance capitalism. Profile Books, London