CULTURES OF EXPERIMENTATION: TESTING INFRA-STRUCTURES IN THE WEB AND BEYOND

By Timo Kaerlein

"Testing – and crucially: experimentally intervening by tweaking the environmental settings – becomes a feature of everyday life when people routinely interact with 'smart' devices and data-intensive media technologies that capture data about their use for constant interpretation and adaptation."

Suggested citation:
Timo Kaerlein, Cultures of Experimentation: Testing Infrastructures in the Web and Beyond. *Interface Critique* 4 (2022): 149–158.
DOI: https://doi.org/10.11588/ic.2023.4.93418

In the summer of 2020, when individuals and governments all over the world were still coming to terms with what would later turn out to have been the first wave of a global pandemic, the web optimization company Optimizely published an e-book titled *Top COVID-19 Experimentation Ideas*. Targeted at businesses it is aiming "to take the lead in the post-pandemic landscape."[1] Since "[l]ockdown measures have driven more people online",[2] Optimizely posits optimistically, the pandemic creates "a once-in-a-lifetime chance for us all to experiment"[3] with changing customer behaviour patterns, new types of users altogether, and a general growth of online communication and consumption. Whereas businesses were still having a hard time figuring out how to adapt to the changing circumstances of digital media use, the authors of the industry guide have a reassuring message for them: "In the middle of all this uncertainty, let's also remember that there are still certainties around which we can reshape our digital strategies."[4]

What are these certainties the digital customer experience experts allude to? Leading business consulting firms all agree, or so the authors want their readers to know, that success in the digital economy can be traced back to a single formula: Experimentation. Testing the performance of different versions of web interfaces against each other on live websites, which might include the tweaking of seemingly insignificant parameters like the precise placement of images and texts or the colour of dialog boxes, has indeed become a standard practice for companies that aim to generate value online. This includes major e-commerce companies, social media platforms, and news websites. To be sure, these practices have not emerged with the pandemic but have been a defining feature of the World Wide Web since the early 2000s. The development of testing infrastructures has gone through a process of professionalisation in the 2010s, with specialised firms employing sophisticated statistical methods and suites of web tools to provide experimentation platforms-as-a-service for web-operating businesses. In today's web environment, two users of a website will rarely see the exact same version of it but will instead be subjected to a never-ending series of tests, adaptations, and performance measurements. Often the language of usability testing is employed to characterise the practice of confronting controlled user segments with slight variations of a website under live conditions. But it is worth inquiring a little deeper into the epistemology, politics, and ethics of web testing infrastructures, not least in order to be able to comprehend their implications for the world in front of the screen.

To this end, I will begin with giving a brief overview on the recent history, contemporary practice, and knowledge claims of digital experimentation plat-

1 Optimizely, Top COVID-19 Experimentation Ideas (2020), p. 3.

2 Ibid., p. 5.

3 Ibid., p. 7.

4 Ibid., p. 4.

forms and testing infrastructures. Having accomplished this, I will then observe a shift in the parameters and subject configurations of these testing regimes that mirrors a broader development in human-computer interaction (HCI) and customer experience design: from cognitive framings of users as goal-oriented rational actors to an understanding of users as suggestible, affect-driven test subjects that can be subtly nudged towards desired action-paths.[5] The conclusion attempts to situate the portrayed development in the light of recent analyses of the ubiquity of testing in computational environments. It also sketches the trajectories of sensor-based testing infrastructures beyond the web.

A/B testing and beyond: On the prevalence of digital experimentation platforms

A/B testing different versions of web interfaces has become a standard web design practice since the late 1990s – and according to one protagonist, "one of the most sacred practices in tech."[6] Big tech companies like Google, Microsoft, Amazon, Facebook, but also more specialised enterprises in the travel and entertainment sector or financial industries engage in "online controlled experimentation"[7] to evaluate the impact of interface design choices on customer behaviour. To this end, live customer traffic is routinely divided into parallel test groups, all the while measuring the impact of different design choices on key-performance indicators (KPIs) like conversion rates or retention times. The overall rationale behind using A/B testing and similar approaches for many companies operating online is an orientation towards data-driven decision-making based on live data from actual user interactions. An "experimentation culture"[8] (as opposed to a mere infrastructure operating in the background) does not only comprise tools and platforms but has quite far-reaching implications for organisational

5 Florian Hadler and Daniel Irrgang, Editorial: Navigating the Human. *Interface Critique Journal* 2 (2018): 7–16.

6 Alex Weinstein, The dark side of A/B testing. *VentureBeat* (April 13, 2019); https://venturebeat.com/2019/04/13/the-dark-side-of-a-b-testing/, access: August 11, 2021, 3:30pm. For a concise overview of the main elements of an A/B testing architecture see Ron Kohavi and Roger Longbotham, Online Controlled Experiments and A/B Testing, in: *Encyclopedia of Machine Learning and Data Mining*, ed. Claude Sammut and Geoffrey I. Webb (New York 2017), pp. 1–8.

7 Aleksander Fabijan, Pavel Dmitriev, Helena Holmstrom Olsson, and Jan Bosch, Online Controlled Experimentation at Scale, in: *Proceedings of the 44th Euromicro Conference on Software Engineering and Advanced Applications* (2018), pp. 68–72.

8 Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin, From Infrastructure to Culture, in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoff Webb, Dragos D. Margineantu, and Graham Williams (New York 2015), pp. 2227–2236, here p. 2227.

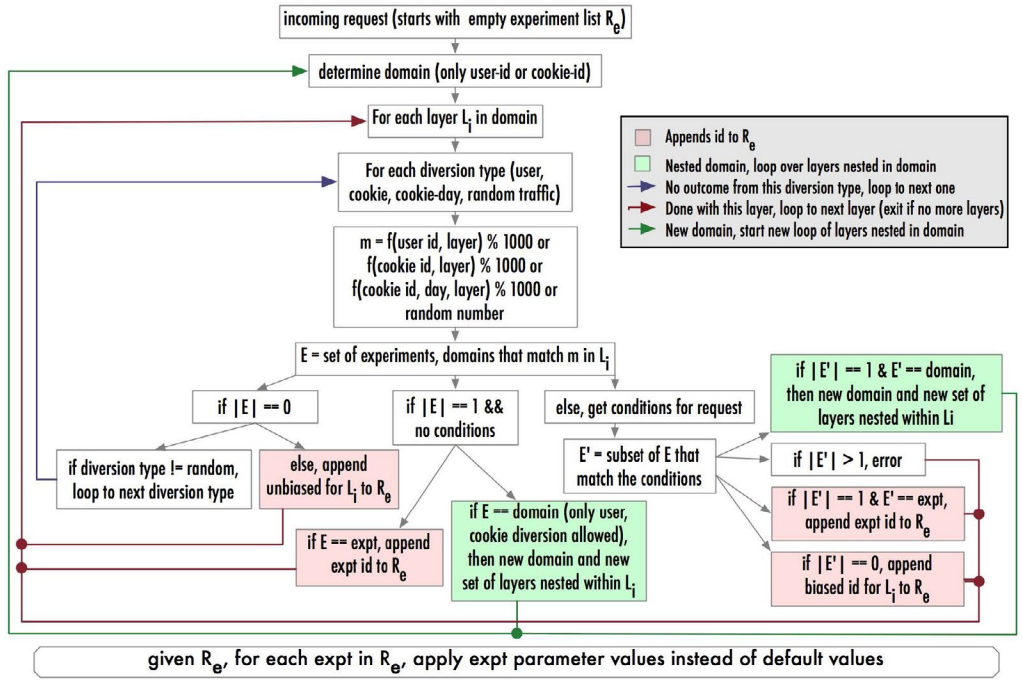

Fig. 1: Logic flow for query requests in Google's overlapping experiment infrastructure. Source: Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer, Overlapping Experiment Infrastructure: More, Better, Faster Experimentation, in: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: ACM Press, 2010), pp. 17–26, here p. 22.

processes, leadership styles and business strategies. Leading authors in the field of web experimentation posit their approach explicitly against outdated HiPPO-based managerial cultures (the acronym stands for "Highest Paid Person's Opinion"),[9] championing instead an evidence-based approach that feeds on large amounts of data.

Major companies often develop in-house experimentation platforms to test the performance not only of visual website elements, but also of different machine learning algorithms like recommendation engines that preselect visible content based on user profiles and preferences. At Google, where "experimentation is practically a mantra,"[10] an overlapping experiment infrastructure has been implemented as early as 2007. The approach builds on already established multi-variate testing schemes that allow for the inclusion of several test factors in parallel,[11] and partitions the various

9 Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne, Controlled Experiments on the Web. *Data Mining and Knowledge Discovery* 18 (2009): 140–181, here p. 178.

10 Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer, Overlapping Experiment Infrastructure, in: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York 2010), pp. 17–26, here p. 17.

11 Kohavi et al., Controlled Experiments on the Web, pp. 158–163.

testing dimensions – e.g., user interface changes, algorithmic variations – into layers of subsets that are designed not to interfere with ongoing experiments in other subsets.[12] Its dimensions are staggering (see fig. 1): At any given point in time, several *billion* possible combinations of test factors are presented to various test groups in parallel, all the while keeping the website's basic functions operational.[13] It is self-evident that no human can make sense of the results of such deeply integrated testing architectures, and the designers readily acknowledge that their scope and flexibility is indeed limited by semantic bottlenecks since it's impossible to understand what exactly is being tested in any given configuration.

Due to the increasing complexity of testing infrastructures in the web (and the increasing demand for fast and reliable data), recent years have seen a process of professionalisation with a range of companies entering the market that offer experimentation platforms-as-a-service, also to medium-sized enterprises. These tie in with existing services of web analytics and search engine optimisation, thus allowing businesses to implement their own individually configured testing architectures. Providers such as Google Optimize, VWO, AB Tasty, and Optimizely develop new statistical methods of continuous monitoring and sequential testing, which make possible, for example, the adjustment of the sample size during a running experiment or the parallel testing of a large number of (computer-generated) hypotheses without the need for human oversight.[14] In web-based experimentation cultures, we can thus observe a *deterritorialisation of the experimental situation* as such, which as a distributed process can no longer be clearly localised and progressively coincides with practices of use.

Who or what is being tested? From usability optimisation to large-scale experiments on users

Not only does the deterritorialisation of the experimental situation refer to the ubiquity of testing practices in web environments (i.e., a matter of scale), but also

12 Tang et al., Overlapping Experiment Infrastructure, pp. 19–21.

13 The paper by Tang et al. doesn't include details on the number of conducted experiments, but Kohavi et al. 2013 report on their work with online based experiments at Microsoft's Bing search engine that references "30 billion possible variants of Bing" in a 2-week testing period. The scale of experimentation at Google is likely to be much higher. See Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, Nils Pohlmann, Online Controlled Experiments at Large Scale, in: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York 2013), pp. 1168–1176, here p. 1168.

14 Leo Pekelis, David Walshy, and Ramesh Johari, The New Stats Engine (Optimizely Whitepaper, undated).

to the scope of their declared purposes and test factors. A particularly notorious example is the widely discussed so-called Facebook emotional contagion study: In January 2012, almost 700,000 Facebook users became unwitting participants in a large-scale experiment to determine the effect on user behaviour of a deliberate manipulation of the number of emotionally positive or negative posts in their respective news streams. Scientists from Cornell University and the University of California, as well as members of the Facebook Research Team, published the results in 2014 and stated the existence of an emotional contagion effect: "[The] results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks."[15] The study quickly sparked controversial discussions because Facebook users did not give informed consent to be included as test subjects, no ethics committee approved its conduct, and risks, such as exposing depressed users to increased negative emotional content, were not considered.[16] Facebook initially maintained that the experiment was essentially nothing more than a usability study, conducted to improve services and provide relevant content to users who had already signed an extensive terms-of-service agreement. The company even went so far as to retroactively update their terms of service to include research as a legitimate scope of internal operations – four months after the controversial experiment had been performed.[17]

The Facebook emotional contagion study has been placed in a direct line of tradition with the Milgram and Stanford Prison psychological experiments, with the crucial difference that the experimental situation of the Facebook case is not framed at all by some laboratory setting but takes place "in the wild" and completely without the knowledge of the participants.[18] While it shares this trait with the majority of experimentation practices in web environments discussed above, it is striking that the purpose of this experiment is decidedly not the improvement of user experiences but the subtle modulation of users' non-conscious affective orientation. Luke Stark has pointed out how the Facebook emotional contagion study but also the large-scale psychographic data profiling based on Facebook data undertaken by Cambridge Analytica in 2016 are rooted in a longstanding "co-development of the psychological and computational scienc-

15 Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111 (2014), pp. 8788–8790, here p. 8788.

16 David Shaw, Facebook's flawed emotion experiment. *Research Ethics* 12/1 (2016): 29–34; Raquel Benbunan-Fich, The ethics of online research with unsuspecting users. *Research Ethics* 13, 3/4 (2017): 200–218.

17 Alex Hern, Facebook T&Cs introduced 'research' policy months after emotion study. *The Guardian* (July 1, 2014); https://www.theguardian.com/technology/2014/jul/01/facebook-data-policy-research-emotion-study, access: August 11, 2021, 5:30pm.

18 Timothy Recuber, From obedience to contagion. *Research Ethics* 12/1 (2016): 44–54.

es."[19] The examples demonstrate how "the clinical psychological subject, a figure amenable to testing and experiment, has been transformed into the scalable subject of social media platforms, structured and categorised by companies like Facebook and universalised as a facet of the lived experience of the digital everyday."[20] With this shift towards the psychometric profiling and micro-targeting of users for economic but increasingly also for political aims, the experimental culture of large-scale testing infrastructures firmly embedded in today's online environments has gained a new urgency and is no longer adequately addressed in terms of usability testing and user experience optimisation.[21]

19 Luke Stark, Algorithmic Psychometrics and the Scalable Subject. *Social Studies of Science* 48/2 (2018): 204–231, here p. 206.

20 Ibid., p. 220f.

21 See Zeynep Tufekci, Engineering the Public. *First Monday* 19/7 (2014) on "real-time, inexpensive and large-scale testing of the effectiveness of persuasion and political communication", already employed in Obama's 2007 presidential campaign. A more technically oriented proof-of-concept for psychometric micro-targeting using Facebook data is elaborated in Till Blesik, Matthias Murawski, Murat Vurucu, and Markus Bick, "Applying big data analytics to psychometric micro-targeting", in: *Machine Learning for Big Data Analysis*, ed. Siddhartha Bhattacharyya, Hrishikesh Bhaumik, Anirban Mukherjee and Sourav De (Berlin, Boston: De Gruyter, 2018), pp. 1–30. It is this journal's declared intention to study interfaces "beyond UX", i.e., to inquire about their history, embedded power relations, and cultural significance. Somewhat ironically, it turns out that interface designers are themselves not primarily "interested in the enhancement of usability, in mere ergonomic questions of design and architecture and in the optimization of user orientation or user experience." (Florian Hadler, Beyond UX. *Interface Critique Journal* 1 (2018): 2–8, here p. 6)

Ubiquitous testing: Sensor-based experimentation in the wild

Furthermore, and this is the last point I would like to argue, the practice of testing and live experimentation on unsuspecting users is currently being extended beyond the borders of the World Wide Web into (mostly urban) public spaces with the help of environmentally embedded sensor media. In line with established notions of ubiquitous computing,[22] the Internet of Things,[23] and 'living lab' approaches in 'Smart City' frameworks,[24] public spaces are increasingly interwoven with semi-autonomous, sensor-equipped devices like 'intelligent' cameras, motion sensors, autonomous cars, drones, and similar technologies. Noortje Marres and David Stark, who also discuss the example of psychographic profiling based on Facebook data, have drawn attention to the circumstance that the epistemology and practices of testing in online environments 'spill over' into the social world at large. They conclude that sociologists need to

22 Mark Weiser, The Computer for the 21st Century. *Scientific American* 265/3 (1991): 94–104.

23 Florian Sprenger and Christoph Engemann (eds.), *Internet der Dinge* (Bielefeld 2015).

24 Jennifer Gabrys, Programming Environments. *Environment and Planning D* 32 (2014): 30–48.

pay more attention to the ways regimes of testing operate not just *in* but *on* social life, i.e., "[w]hereas we traditionally think about testing taking place *within a setting*, today's engineers are *testing the settings*."[25] While in traditional field tests the prior existence of a field is presupposed, the types of technology-intensive testing increasingly encountered today create their own test environments by working through and acting upon social environments. Testing – and crucially: experimentally intervening by tweaking the environmental settings – becomes a feature of everyday life when people routinely interact with 'smart' devices and data-intensive media technologies that capture data about their use for constant interpretation and adaptation. It stands to reason that the established cultures of experimentation in web-based environments outlined above act as a model and inspiration for the plethora of practices of testing and live experimentation witnessable in data-infused real-world environments, not the least because many of the major commercial actors are active in both domains. In a 2012 *Wired* article on the state of the art of A/B testing in web design, author Brian Christian speculated on its prospects of being applied to the physical reality outside the web: "Many web workers, having tasted of the A/B apple, can no longer imagine operating in any other environment. Indeed, they begin to look with pity on the offline world, a terrifying place where each of us possesses only one life to live rather than two (or more) in parallel."[26] Ten years on, the ubiquity of real-time testing and experimentation in data-saturated environments has become an integral element of digital cultures – and its implications for the conduct of everyday life are just beginning to unravel.

25 Noortje Marres and David Stark, Put to the Test. *The British Journal of Sociology* 71 (2020): 423–443, here p. 435.

26 Brian Christian, The A/B Test. *Wired* (April 24, 2012); https://www.wired.com/2012/04/ff-abtesting/, access: August 11, 2021, 6:00pm.

References

Benbunan-Fich, Raquel, The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation. *Research Ethics* 13/3-4 (2017): 200–218. DOI: 10.1177/1747016116680664.

Blesik, Till, Matthias Murawski, Murat Vurucu, and Markus Bick, "Applying big data analytics to psychometric micro-targeting", in: *Machine Learning for Big Data Analysis*, ed. Siddhartha Bhattacharyya, Hrishikesh Bhaumik, Anirban Mukherjee and Sourav De (Berlin, Boston: De Gruyter, 2018), pp. 1-30.

Christian, Brian, The A/B Test: Inside the Technology That's Changing the Rules of Business. *Wired* (April 24, 2012); https://www.wired.com/2012/04/ff-abtesting/, access: August 11, 2021, 6:00pm.

Fabijan, Aleksander, Pavel Dmitriev, Helena Holmstrom Olsson, and Jan Bosch, Online Controlled Experimentation at Scale: An Empirical Survey on the Current State of A/B Testing, in: *Proceedings of the 44th Euromicro Conference on Software Engineering and Advanced Applications* (2018), pp. 68–72. DOI: 10.1109/SEAA.2018.00021.

Gabrys, Jennifer, Programming Environments: Environmentality and Citizen Sensing in the Smart City. *Environment and Planning D: Society and Space* 32 (2014): 30–48. DOI: 10.1068/d16812.

Hadler, Florian, Beyond UX. *Interface Critique Journal* 1 (2018): 2–8. DOI: 10.11588/ic.2018.0.45695.

Hadler, F. and Daniel Irrgang, Editorial: Navigating the Human. *Interface Critique Journal* 2 (2018): 7–16. DOI: 10.11588/ic.2019.2.67261.

Hern, Alex, Facebook T&Cs introduced 'research' policy months after emotion study. *The Guardian* (July 1, 2014); https://www.theguardian.com/technology/2014/jul/01/facebook-data-policy-research-emotion-study, access: August 11, 2021, 5:30pm.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock, Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111 (2014), pp. 8788–8790.

Kohavi, Ron, Roger Longbotham, Dan Sommerfield, and Randal M. Henne, Controlled Experiments on the Web: Survey and Practical Guide. *Data Mining and Knowledge Discovery* 18 (2009): 40–181. DOI 10.1007/s10618-008-0114-1.

Kohavi, R., Alex Deng, Brian Frasca, Toby Walker, Ya Xu, Nils Pohlmann, Online Controlled Experiments at Large Scale, in: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: ACM Press, 2013), pp. 1168–1176.

Kohavi, R. and Roger Longbotham, Online Controlled Experiments and A/B Testing, in: *Encyclopedia of Machine Learning and Data Mining*, ed. Claude Sammut and Geoffrey I. Webb (New York: Springer Science+Business, 2017), pp. 1–8. DOI: 10.1007/978-1-4899-7502-7 891-1.

Marres, Noortje and David Stark, Put to the Test: For a New Sociology of Testing. *The British Journal of Sociology* 71 (2020): 423–443. DOI: 10.1111/1468-4446.12746.

Optimizely, Top COVID-19 Experimentation Ideas (2020).

Pekelis, Leo, David Walshy, and Ramesh Johari, The New Stats Engine (Optimizely Whitepaper, undated).

Recuber, Timothy, From obedience to contagion: Discourses of power in Milgram, Zimbardo, and the Facebook experiment. *Research Ethics* 12/1 (2016): 44–54. DOI: 10.1177/1747016115579533.

Shaw, David, Facebook's flawed emotion experiment: Antisocial research on social network users. *Research Ethics* 12/1 (2016): 29–34. DOI: 10.1177/1747016115579535.

Sprenger, Florian and Christoph Engemann (eds.), *Internet der Dinge. Über smarte Objekte, intelligente Umgebungen und die technische Durchdringung der Welt* (Bielefeld: Transcript, 2015).

Stark, Luke, Algorithmic Psychometrics and the Scalable Subject. *Social Studies of Science* 48/2 (2018): 204–231. DOI: 10.1177/0306312718772094.

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer, Overlapping Experiment Infrastructure: More, Better, Faster Experimentation, in: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: ACM Press, 2010), pp. 17–26.

Tufekci, Zeynep, Engineering the Public: Big Data, Surveillance and Computational Politics. *First Monday* 19/7 (2014). DOI: https://doi.org/10.5210/fm.v19i7.4901.

Weinstein, Alex, The dark side of A/B testing. *VentureBeat* (April 13, 2019); https://venturebeat.com/2019/04/13/the-dark-side-of-a-b-testing/, access: August 11, 2021, 3:30pm.

Weiser, Mark, The Computer for the 21st Century. *Scientific American* 265/3 (1991): 94–104.

Xu, Ya, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin, From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks, in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoff Webb, Dragos D. Margineantu, and Graham Williams (New York: ACM Press, 2015), pp. 2227–2236.