Embracing Data Noise

Ida Larsen-Ledet
Microsoft Research Cambridge
Cambridge, United Kingdom
t-idal@microsoft.com

KEYWORDS

data, noise, design, artificial intelligence

"The skill of data storytelling is *removing the noise* and focusing people's attention on the key insights." — Brent Dykes, supposedly (according to Tracy Mayor of MIT Sloan [14])

"[…] I try to follow the threads where they lead in order to track them and find their tangles and patterns crucial *for staying with the trouble* in real and particular places and times." — Donna J. Haraway, Staying with the Trouble: Making Kin in the Chthulucene [10]

In one way or another, we probably all recognize the discord one is faced with when people's reality must be represented for a computational system to deal with it: Computer systems operate with explicit definitions and discrete structures, but human existence does not lend itself to strict boundaries or clear definitions (as evidenced by millennia of philosophy). I am currently experiencing this tension first-hand in my work with organizational knowledge bases that use machine learning to automatically interpret, organize, and summarize the information that large organizations have in their IT systems [23]. On the one hand, we want a system that is faithful and helpful to the ways of knowing [16] at work in a given organization; on the other hand, this involves breaking information and ways of knowing into (computational) bits, and any such categorization necessarily involves abstraction [4] and so cannot allow 100 % fidelity. And, ideally, the categorization should be futureproof: the system needs to be able to absorb direct and indirect input from organizational knowledge that develops and expands in step with living organizational work and practices. To guide the system on how to make the interpretations necessary for this, we are making choices that involve noise: Put somewhat simplistically, what input will we be expecting and what remaining potential input will, consequently, be noise that lies outside what we plan for the system to produce useful output from? These choices involve determining what boundaries to draw and where to draw them: Is a free-text note valid input? Is an emoji? And what are the consequences of deciding that something is "invalid"?

This conundrum involves the conceptualization of signal [1] versus noise: What user input is meaningful and what is nonsense (to the system)? For instance, it may be argued that accepting free-text input would deteriorate the strength of a system that relies on structured information. But is it fair to require users to change their way of expressing things so that the system will not ignore them? Aren't we making the system for the benefit of the users? Bowker and Star's work clearly showcases the challenges and, oftentimes negative, impact of representing the world using rigidly defined categories [4].

This has prompted the question for me: How can systems be designed and created with and for noise? Below, I discuss examples that involve conceptualization, acceptance, and use of noise; including what may be gained from viewing seemingly undesirable output as noise with potential. I end with some brief reflections on what it could mean to embrace noise when designing computer systems.

1 SOME CONCEPTUALIZATIONS OF NOISE

Noise is a helpful concept in fields like data science, machine learning, and artificial intelligence (AI). It can help make data manageable, for example by allowing "noisy" data points to be removed so the data can be streamlined to fit a computational structure. If everything is accommodated somewhere, nice categories will become polluted with things that don't actually belong (see also Bowker and Star on residual categories [4]). There's a reason many people have a drawer or box somewhere for random knick-knacks that don't fit anywhere [18] – at least all the other drawers will be neat and tidy (and, of course, you'll only have to look in one place for things that don't have an obvious category). **Noise can be a way to name what we are unable to categorize.**

Noise can also come up when conditions are not ideal. With computers, there are usually constraints on disk space or processing power. Less-than-ideal conditions introduce the need for trade-offs. In the case of portable MP3 players, for instance, a desire for 1-1 representation took a backseat in favor of increasing the number of songs that could be kept on a device, with the help of audio compression. This introduced some distortion – noise – but gave us the ability to carry a library of music in a pocket. **This concept of noise is one of undesirable, but potentially acceptable, artifacts that detract from what we're actually interested in.**

Today's diffusion models take the presence of noise a step further, not only accepting noise but actually making use of it. Diffusion models work by synthesizing realistic things from random noise [2]. The models are trained to do this by recreating clean input from noisy versions of that input: The input is distorted through the introduction of noise – visually, you can picture this as transforming a sharp image to look like the white static on an old television set that isn't receiving a signal. The training then consists in returning the distorted input back to its original state – or *denoising* it. The introduction of noise happens by replacing the values of individual data points with new values drawn from a statistical distribution around the original data points. The denoising process relies on, or can at least benefit from, knowing the statistical distribution that the values were drawn from – i.e., using a specific distribution gives

[1] Essentially, signal in this context refers to data that somehow carries meaningful or useful information – as opposed to data that is noise.

[2] A helpful introduction is provided in this YouTube video by AssemblyAI: https://www.youtube.com/watch?v=yTAMrHVG1ew. The following two blog are also helpful for understanding the basics: https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/ and https://towardsdatascience.com/stable-diffusion-best-open-source-version-of-dall-e-2-ebcdf1cb64bc.

us control over the situation by letting us inform the model of what distribution the noise came from. Using a well-defined probability distribution also means that the model can be optimized to be faster to train, and thus more practical, by relying on probabilistic properties of the distribution; for example, knowing that applying multiple Gaussian distributions consecutively corresponds to applying one Gaussian with the combined parameters. **The concept of noise applied here is one of controlled and predictable interference that is, eventually, removed.**

2 THE NUANCES OF NOISE

These examples show why we may sometimes need to talk about noise, and why we may want to get rid of it. But they also show that we miss out if we reduce the concept of noise (or clutter [19]) to something that must be eliminated at all costs (see also Taylor et al. [19] for a critique of viewing the diversity and inconsistencies of human existence as problematic) [3]. Yet, the fuzziness of the human condition is still often treated as something to be smoothed out before computational systems can be bothered to deal with it [17]. The dents and crevices of human ways of being and acting are framed as bothersome. Anything that is not what a given system set out to do is noise to that system, requiring people to segment their existence across different technologies and adapt to their limitations [2, 3, 12]. By framing human nuance as noise that is not worth dealing with, we are asking people to accommodate to technology rather than the other way around.

The way the term "noise" is currently used, it can comprise **anything from truly irrelevant signals to unexpected, yet very important, things.** As hinted at above, there is an important difference between noise in the sense of, for example, an audio recording or an image, where the original auditory or visual impression has been distorted, and noise in the sense of things that we can't get to fit our attempt at discretizing reality. Some would argue that something qualifies as noise if there is no pattern to it – not least in the context of machine learning. But this would mean that the definition of noise changes depending on the scale we look at: A pattern may be possible to discern at one scale but not at another [13] – see Figure 1.

Those who define noise as lack of regularity might well acknowledge the role played by scale of perspective. But in that case, can we change the scale so there's room for humans to be human [13]? Where would that change need to happen? Is it about how we look at the world or about how fine-grained a digital representation we can devise? The latter sets a goal of recasting apparent noise as signal, whereas the former can still require us to accept noise as a condition.

3 HALLUCINATIONS, APPREHENSION, AND POTENTIAL

To further examine the apprehension we can feel towards noise, let's look to a recent example: With the hype, curiosity, and fear surrounding the release of Stable Diffusion, ChatGPT, and other AI models unlike anything the public has experienced before has also

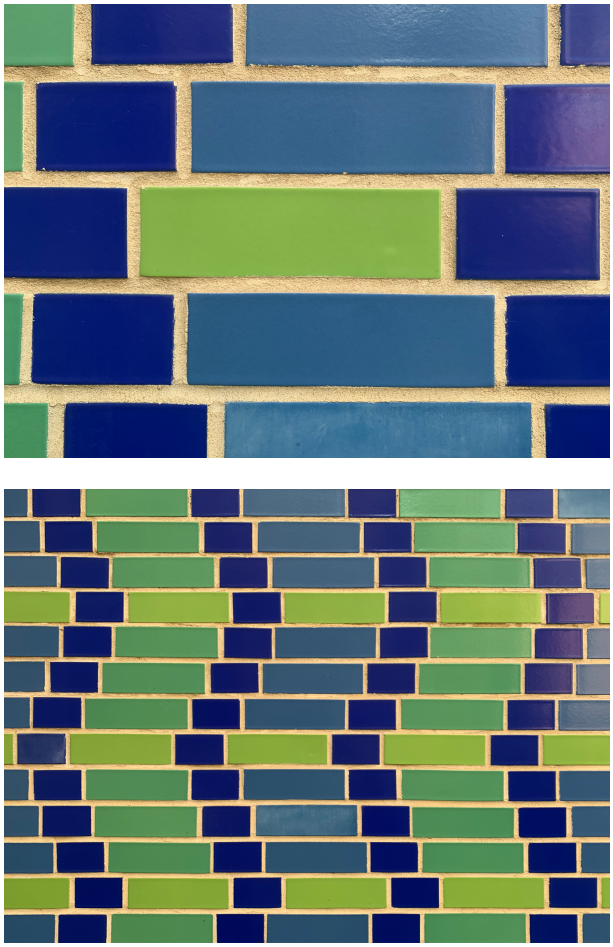

Figure 1: The green brick might seem out of place in the close-up but not in the photo taken from further away.

come the propensity for these models to *hallucinate* [5]. Hallucination, in this context, refers to the generation of output (usually text) that contains falsities. These hallucinations pollute the intended factual – and, purportedly, thus also reliable and trustworthy – service. In doing so, they are a form of noise.

But could an AI hallucination be used in a controlled way? Could it give us something unexpected but important? Could it be valuable or useful? For decades, people have dreamt of creative AI, and both experts and laypeople have argued on multiple occasions that true intelligence is creative. Stifling the, for all intents and purposes, creative tendencies of these models would seem to halt the realization of this vision in its infancy [4]. And can we even know that we will be able to remove or suppress AI hallucinations? We may instead accept that this noise is part of the package. This would present a great challenge to the designers who will craft the experiences people can have with AI models. So far, not much design has gone into those experiences – they are currently limited to a text input field, a few buttons, and spaces to display what the model

[3] In the book *Effective Data Storytelling*, Brent Dykes talks about removing *unnecessary* noise [6]; there's an implicit acknowledgement that some noise may be worth keeping in order to paint the right picture.

[4] Whether that would, in fact, be such a shame is no doubt a highly contested matter.

has produced in response to input. But once these experiences become more designed, a lot of power and responsibility will lie with designers. And working out how to design with noise could be a way to tackle the trouble of hallucinations, perhaps by working with them.

Let's discuss AI hallucinations as noise with potential. While the relationship between AI hallucinations and human hallucination is only surface-level, I am compelled to draw parallels to the role played by psychedelics in the work of known artists [22], and to a recent Nature video that compared deliberately inducing states of altered consciousness (such as hallucinations) to "throwing a stone into a still pond to see what happens" [15]. The act of throwing a stone into a pond is not necessarily harmless but not inherently bad either. Although machine learning scientist Catherine Breslin's wording when describing AI hallucinations is unfaltering: "LLMs can hallucinate facts and generate text which is just wrong," [5]. I would argue that "wrong" is in the eye of the beholder – untrue facts are wrong in a textbook but can be exactly right in a novel. Some writers are already using AI hallucinations to unsettle the pond and stimulate their creativity [7].

Taking things to quite a daunting level, neuroscientist Chris Timmermann explains that altered consciousness may allow people to "access some of the brain mechanisms that allow us to construct realities" [15]. The reason AI hallucinations are frightening is that they, similarly, take part in constructing (or reconstructing [8]) our reality through the impressions they cause us to form. Perhaps the horror is in the very fact that *generated* does not necessarily equal *fake*, and that the question of whether something depicts reality or not can be a muddled one – as when AI hallucinations are used to recreate occluded parts of an image [8]: To what extent are we seeing an accurate reconstruction of the truth, and to what extent is it made up?

And all in all, is the danger really in the hallucination or in the fact that we're not sure what is hallucination and what is reality? And is it truly danger, or is it just as much unease at treading into an uncanny valley where we're faced with an ambiguity of fact versus fiction that we have not previously had to deal with in machines?

4 FINDING TRUTH IN THE GREY AREAS

Ambiguity can be unpleasant – but also liberating. Understanding how to use ambiguity in design [9] could prove effective in helping us navigate a noisy world, if we cannot or do not want to get rid of the noise. The essence of ambiguity, as Gaver et al. [9] talk about it, is *openness to interpretation*: They argue that ambiguity encourages people to interpret situations for themselves and "[grapple] conceptually with systems and their contexts" [9, p. 233]. There is a form of dialectic relationship between ambiguity and noise, wherein noise can produce ambiguity, and ambiguity can give an impression of noise (think about how things end up in the knick-knack drawer). The ambiguity created by noise helps us put into question what might otherwise have been presented or perceived as fact. The notion of "facts" tends to draw us toward a logical understanding of true or false (well-suited for computers). If we apply this understanding uncritically, the logic of factuality spills over into areas where it can misguide us. For example, an organizational knowledge base may – correctly – contain the fact that an organization uses a particular process for part of their operation. But the process as such is not a fact – yet, neither is it a lie. The process exists in action and so is, effectively, a fiction as long as it is not being put into existence through execution – and each and every execution will be different from the others. Grey areas are where such things live: These things that are not fact, but which are not untrue either. Factuality cannot capture this. Seen through a factuality lens, the different executions of the process appear noisy because they don't provide a consistent "answer"; yet they are more authentic than their abstract representation.

Removing noise can involve removing nuances that could help people interpret things. Without the "noise" of the world, people get a reduced picture [20]. Take, again, the organizational knowledge base: Imagine letting people in the organization annotate knowledge base entries with free-text. This is a nightmare in terms of structured data and categorization, as people might make the same annotation in different ways, use the same word to mean different things, or not agree on what sorts of things are relevant to mention. But for someone looking up that entry, seeing that it is full of annotations from many different people might tell them that the topic of the entry is used in many different contexts or in many different, potentially incongruent, ways [1]. Cleaning this up would make the entry appear less contested than it is. Depending on the situation, this could be harmless or detrimentally misleading. There is an element of respect for people in embracing noise by drawing attention to the traces of people's practice without dictating how to interpret those traces (similarly to what Gaver et al. have said about ambiguity [9]).

5 DESIGNING WITH NOISE

Just as Gaver et al. framed ambiguity as an opportunity rather than a problem for design, noise could be a resource rather than a nuisance – or, at least, something to work *with* rather than against. In the same way that ambiguity leaves room for people to make sense of things in their context and can "lead to a deep conceptual appropriation of the artifact" [9, p. 236], leaving room for noise could produce knowledge bases that are better adopted into people's practices. Not necessarily more easily adopted – users will need to put in effort – but better adopted.

I am not arguing noise should be uncritically embraced. Removing or reducing noise can be the right choice – just as lack of ambiguity can be, e.g., in safety-critical environments [9]. But neither should noise be uncritically disregarded. If we teach ourselves to look at noise in *a nuanced way*, we may find ourselves better able to apply it in *useful ways*. In the above, I've presented a couple of takes on noise: The uncategorizable; undesirable artifacts; controlled interference; irrelevant signals; unexpected but valuable data points; falsities; pollution; unstructured data. In talking about these different kinds of noise, I have also touched on potential ways of working with them. I hope this will be a seed for future work to develop a sophisticated perspective on noise and its challenges and utilities in design for human lives and practices.

We could also accept elimination of noise and use ambiguity as a tool to remedy this[5]. We may continue to eliminate noise but

[5] See [24] for a technical analysis of the relationship between AIs and expressions of uncertainty, including a brief discussion of opportunities and risks associated with training AIs to understand and express uncertainty.

present these systems more honestly – acknowledging that "something is missing" – to encourage users to embrace the system as a partial truth, by "[impelling] people to question for themselves the truth of a situation" [9, p. 240]. Part of this acceptance would involve recognizing that representations do not have to be 1-1 or even approach that close of a match. In many cases, the value of representation lies in abstraction: If the representation provides nothing in the form of re-mediation, we may as well be interacting with the real thing instead of the representation. The value of *re*-presentation is to show things in a different way, to highlight, to emphasize, to filter, and/or to contain something in a different medium – for instance, "containing" part of the world in a knowledge base or preserving audio digitally to be able to listen to a piece of music whenever we please.

It's also possible that new developments in machine learning and AI that rely less on human guidance, such as foundation models, will be able to work with noise and perhaps even draw utility from it. The challenge for design might then be one of enabling interpretation and interrogation of the system's reasoning [21] and helping users understand how to influence the system [see e.g., 11] – perhaps to help people work out how to make the right noises to get the system to do what they need.

Human existence is noisy. Let's embrace that.

ACKNOWLEDGMENTS

Thanks to Pratik Ghosh for inspiration, to Max Croci for sense-checking and answering my questions about diffusion models, and to Siân Lindley for feedback.

REFERENCES

[1] Mirzel Avdic, Susanne Bødker, and Ida Larsen-Ledet. 2021. Two Cases for Traces: A Theoretical Framing of Mediated Joint Activity. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 190 (apr 2021), 28 pages. https://doi.org/10.1145/3449289

[2] Liam Bannon, Allen Cypher, Steven Greenspan, and Melissa L. Monty. 1983. Evaluation and Analysis of Users' Activity Organization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '83)*. Association for Computing Machinery, New York, NY, USA, 54–57. https://doi.org/10.1145/800045.801580

[3] Jakob E. Bardram, Steven Jeuris, Paolo Tell, Steven Houben, and Stephen Voida. 2019. Activity-Centric Computing Systems. *Commun. ACM* 62, 8 (jul 2019), 72–81. https://doi.org/10.1145/3325901

[4] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences.* The MIT Press.

[5] Catherine Breslin. 2022. What are Large Language Models? A look at LLMs and their popularity. Retrieved February 17, 2023 from https://chatbotslife.com/what-are-large-language-models-7b3c4c15e567

[6] Brent Dykes. 2019. *Effective data storytelling - how to drive change with data, narrative and visuals.* Wiley.

[7] Josh Dzieza. 2022. The Great Fiction of AI — The strange world of high-speed semi-automated genre fiction. Retrieved February 17, 2023 from https://www.theverge.com/c/23194235/ai-fiction-writing-amazon-kindle-sudowrite-jasper (The Verge).

[8] Federico Fulgeri, Matteo Fabbri, Stefano Alletto, Simone Calderara, and Rita Cucchiara. 2019. Can Adversarial Networks Hallucinate Occluded People With a Plausible Aspect? *CoRR* abs/1901.08097 (2019). arXiv:1901.08097 http://arxiv.org/abs/1901.08097

[9] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a Resource for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/642611.642653

[10] Donna J. Haraway. 2016. *Staying with the Trouble: Making Kin in the Chthulucene.* Duke University Press. http://www.jstor.org/stable/j.ctv11cw25q

[11] Wendy Ju and Larry Leifer. 2008. The design of implicit interactions: Making interactive systems less obnoxious. *Design Issues* 24, 3 (2008), 72–84.

[12] Heekyoung Jung, Erik Stolterman, Will Ryan, Tonya Thompson, and Marty Siegel. 2008. Toward a Framework for Ecologies of Artifacts: How Are Digital Artifacts Interconnected within a Personal Life?. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges* (Lund, Sweden) *(NordiCHI '08)*. Association for Computing Machinery, New York, NY, USA, 201–210. https://doi.org/10.1145/1463160.1463182

[13] Ida Larsen-Ledet, Ann Light, Airi Lampinen, Joanna Saad-Sulonen, Katie Berns, Negar Khojasteh, and Chiara Rossitto. 2022. (Un) Scaling Computing. *Interactions* 29, 5 (aug 2022), 72–77. https://doi.org/10.1145/3554926

[14] Tracy Mayor. 2021. 15 quotes and stats to help boost your data and analytics savvy. Retrieved February 23, 2023 from https://mitsloan.mit.edu/ideas-made-to-matter/15-quotes-and-stats-to-help-boost-your-data-and-analytics-savvy (MIT Sloan).

[15] Nature video (YouTube channel). 2022. Psychedelics and consciousness: Could drugs help quantify our waking state? Retrieved February 17, 2023 from https://www.youtube.com/watch?v=1Tl7ktZLWF8

[16] Wanda J. Orlikowski. 2002. Knowing in Practice: Enacting a Collective Capability in Distributed Organizing. *Organization Science* 13, 3 (2002), 249–273. http://www.jstor.org/stable/3086020

[17] Lucy A. Suchman and Randall H. Trigg. 1993. *Artificial intelligence as craftwork.* Cambridge University Press, 144–178. https://doi.org/10.1017/CBO9780511625510.007

[18] Laurel Swan, Alex S. Taylor, Shahram Izadi, and Richard Harper. 2007. Containing Family Clutter. In *Home Informatics and Telematics: ICT for The Next Billion.* Springer US, Boston, MA, 171–184.

[19] Alex S. Taylor, Richard Harper, Laurel Swan, Shahram Izadi, Abigail Sellen, and Mark Perry. 2007. Homes that make us smart. *Personal and Ubiquitous Computing* 11 (2007), 383–393. https://doi.org/10.1007/s00779-006-0076-5

[20] Peter Tolmie, Andy Crabtree, Tom Rodden, James Colley, and Ewa Luger. 2016. "This Has to Be the Cats": Personal Data Legibility in Networked Sensing Systems. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 491–502. https://doi.org/10.1145/2818048.2819992

[21] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[22] Holly Williams. 2018. How LSD influenced Western culture. Retrieved February 25, 2023 from https://www.bbc.com/culture/article/20181016-how-lsd-influenced-western-culture (BBC).

[23] John Winn, Matteo Venanzi, Tom Minka, Ivan Korostelev, John Guiver, Elena Pochernina, Pavel Myshkov, Alex Spengler, Denise Wilkins, Siân Lindley, Richard Banks, Sam Webster, and Yordan Zaykov. 2021. Enterprise Alexandria: Online High-Precision Enterprise Knowledge Base Construction with Typed Entities. In *Automated Knowledge Base Construction.* https://www.microsoft.com/en-us/research/publication/enterprise-alexandria-online-high-precision-enterprise-knowledge-base-construction-with-typed-entities/

[24] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the Grey Area: Expressions of Overconfidence and Uncertainty in Language Models. https://doi.org/10.48550/ARXIV.2302.13439