GenAI, the “Chain of Accountability”, and gen-ethics.

by Katja RAUSCH

With 100 million users two month after launch in January 2023, ChatGPT may be the fastest-growing consumer internet app ever, as reported by the Guardian. An unprecedented phenomenon for a large language transformer model. And a new technological terminology becoming mainstream overnight: generative AI or genAI.

However after the initial frenzy, and the bot's hallucinating outputs, a sobering question quickly became pivotal: What about responsibility and generative AI?

Rhetorical questions triggered abundantly: Do we want responsible innovation? Do we want responsible AI? Do we want responsible genAI?

Of course, we do! Then why does ethics have such a hard time with genAI?

Recent events, such as Microsoft laying off its ethics team, or statements by Sam Altman, CEO at OpenAI on how “to bake in ethics” have painfully and shockingly revealed how low ethics, responsibility and accountability are viewed by major tech players.

> *When we expect the highest LEVEL of responsibility, we see the lowest SENSE of responsibility.*

Nihil novi sub sole? Maybe yes. What really does change is that genAI is a high-speed planetary phenomenon; beyond technology. It has been adopted instantly by people (even though it is going rogue), and has already infiltrated numerous existing systems (by law firms such as Allen & Overy "to boost legal work, hospitals, finance, logistics providers...) at the speed of light.

It is a systemic phenomenon integrating technological systems, business processes, social systems, human systems, thus value systems.

Yet from an ethical side, genAI sits on ethical permafrost. By now, we all know, and it remains uncontested, that large language models are based on unprecedented industry-scale Intellectual Property thefts, data privacy violations and copyright infringements.

All hidden by a playful, apparently harmless user interface with an actual potential to be a superspreader of misinformation, and a fountain of hostility, intimidation and nonsense.

Let's hear it from the horse's mouth:

> *“However, it is important to note that chatbots are still a relatively new technology and there are potential challenges and risks associated with their use, such as the risk of bias or inaccuracies in their responses.“ (Source: OpenAI)*

Sam Altman, CEO OpenAI

Or Mira Murati, CTO, OpenAI on ChatGPT in the Time article "The Creator of ChatGPT Thinks AI Should Be Regulated".

> *"ChatGPT is essentially a large conversational model - a big neural net that’s been trained to predict the next word - and the challenges with it are similar challenges we see with the base large language models: it may make up facts.*

Is this the price we have to pay for innovation? Should we use the usual risk-benefit reasoning for guiding our ethical decisions? Like national governments did for COVID-19 vaccinations?

Well that's one of the arguments used by the developers (OpenAI), their allies (Microsoft), and even competitors. A self-serving *consequentialist ethics* reasoning (will be debated later).

Fact is that Pandora's Box is wide open, and constantly generating new questions.

The genAI integration race across industries is turning at full speed. ChatGPT, Bing, Bard or Ernie or any newly developed conversational transformer model will no longer be stand-alones. They will be plugged in, integrated or patched on existing business processes as facilitators, co-pilots, assistants, organizers, creators, advisors, accelerators...

On March 23rd, OpenAI released an additional set of plugins for ChatGPT with the notice

> *"In line with our iterative deployment philosophy, we are gradually rolling out plugins in ChatGPT so we can study their real-world use, impact, and safety and alignment challenges - all of which we'll have to get right in order to achieve our mission.*
>
> *Users have been asking for plugins since we launched ChatGPT (and many developers are experimenting with similar ideas) because they unlock a vast range of possible use cases."*

OpenAI's mission being "planning for AGI and beyond". The ultimate race. The Graal.

However, with high-speed genAI industry integrations, companies must rethink their entire **Chain of Accountability** before being hit by upcoming regulation.

Recently, the American Federal Trade Commission (FTC), after helping for years to build the current monopolistic IT landscape in the US, officially declared to 'bust the trust' of today's digital fiefdoms, and begin forging a new social contract for the Internet era." under Chair Lina Kahn. source

Today in Europe, the most developed proposal for regulating AI comes from the European Union, which first issued its Artificial Intelligence Act in 2021. The final version, probably integrating a specific section on genAI, is about to be released.

Consequently a general-purpose generative AI system like ChatGPT or Bing or Bard, used by millions of people could fall under the "high risk" category, and consequently be banned in the EU. But we all know it won't... There are other options, namely shift focus.

> *With a shifted focus on HOW technology is used, its purpose, rather than technology itself, the notion of accountability is of highest importance.*

For executives, in private and public sectors alike, crucial questions will be: How to deal with genAI? How to integrate it in our existing business model?

To address this question, two main angles will be scrutinized: first, the need for a "Chain of Accountability", and second, the need for a new kind of gen-ethics.

I. The Chain of Accountability

The following reflection is based on numerous years as an IT consultant in NY and Paris, and teaching Information Systems in Logistics at the Sorbonne for 12 years.

The reference to logistics, and consequently supply chain management can provide excellent guidance for companies to manage upcoming challenges while integrating and using genAI.

The business model

When considering genAI beyond the technological product it is, one needs to consider its business model.The current generative AI business model has transitioned from an initially direct business model (OpenAI ChatGPT to end user) into a transformative intermediary model (OpenAI ChatGPT via Microsoft - or any other company - to end user). This adds several new layers of complexity. From a technological side for sure, but more so from ethical and legal sides. We focus on the ethical side.

In an intermediary model most companies will integrate large language modules or plugins (ChatGPT or others extensions) into their existing business architecture. By doing so there is an urging need to widen the perspective.

Like logistics providers, integrative companies need to privilege a holistic view: they need to consider the upstream and downstream implications, not only from a technological side but mostly from a regulatory and ethical side.

Producing and integrating companies, in an intermediary business model, always act as the “middleman”. There is a before and an after on the Value Chain, on the Chain of Production, on the Chain of Communication. Thus on the “Chain of Accountability”.

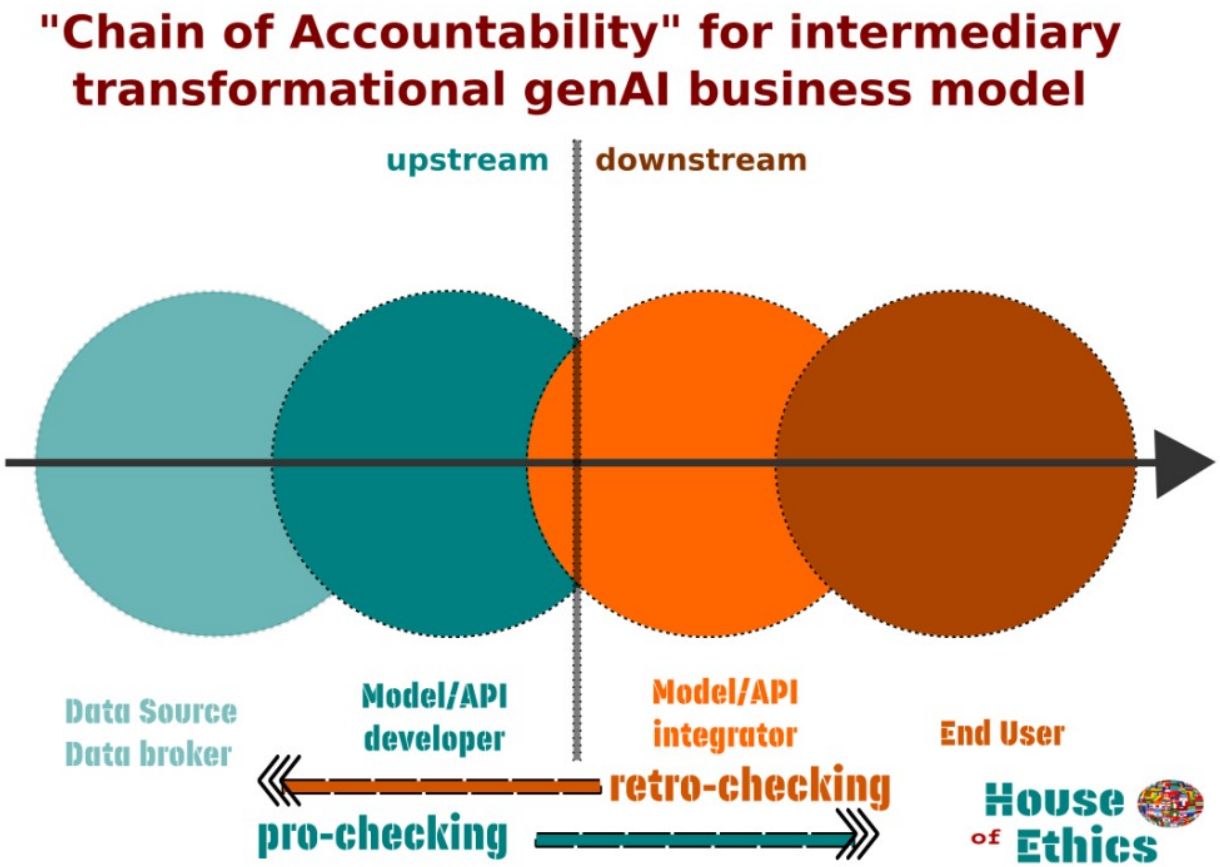
The dynamics of the “Chain of Accountability”

Any supply chain comes with tier 1, tier 2 or tier 3 up- and downstream suppliers. A chain of reaction, triggered by a flow of information, data, and products where forecasting and retro-planning is paramount. Quality management, risk assessment, and data governance are daily bread.

To adopt this holistic approach, a logistician’s up-and downstream view, in managing ethical matters and dealing with accountability will become a distinctive advantage for integrators of genAI tools.

Only those companies viewing genAI as a process, and not as a tool, will be fit for the challenge. Will be capable of aligning multi-layered imperatives, and generate success.

They will avoid the lurking pitfalls of a tunnel vision and the bullwhip effect in accountability and responsibility.



How does the "Chain of Accountability" work?

To illustrate the process, we have used the popular OpenAi, ChatGPT, Microsoft, Bing constellation.

Upstream

- **Level 1: Data brokers, Data source**

In the case of ChatGPT, Common Crawl provides 60% of ChatGPT's training data. This data is based on scraped internet data: people data! “Found” on the internet and used without consent.

Commen Crawl: "We didn’t produce the crawled content, we just found it on the web. So we are not vouching for the content or liable if there is something wrong with it." (source: ToU Common Crawl)

Data used for Meta's LLaMA has been similarly trained on massive data being scraped.

In this case, responsibility and accountability is clearly delegated to the next level. In a direct business model, the next level being the end user. In an intermediary model, the next accountable level is the model developer or producer.

The duties of the producer reside in retro-checking and pro-checking accountability. Is this data good for my model? Do I respect the end user with my service or product?

- **Level 2: the producing/developing company**

Mira Murati: “It’s important for OpenAI and companies like ours to bring this into the public consciousness in a way that’s controlled and responsible. But we’re a small group of people and we need a ton more input in this system and a lot more input that goes beyond the technologies - definitely regulators and governments and everyone else.”

In this case, OpenAI, retro-actively has used data which is not collected in a responsible manner, and delegates responsibility one level down to the data broker; simultaneously they also delegate a major part of accountability one level up to the end user, and even to outside supervising institutions like regulators or government. Just lately Altman mentioned "the entire society" to figure out the problem.

We all understand that OpenAI's duty in this matter is not to "bring it to the public's consciousness" but to put a robust, transparent, trustworthy model on the market ready for responsible and safe use.

Such "cart before the horse" behavior might be scrutinized by regulators in the near future.

Delegating harmful outcomes to end users might be a risky choice of action for the near future. Companies better plan on a safe and solid *Chain of Accountability*.

Downstream

- **Level 3: the integrating company**

This level will be the most common and of highest interest for developing a "Chain of Accountability" plan.

The integrating company has to pro-check and retro-check. Integration should have a Two-Factor Authentication (2FA) for Accountability, similar to the data protection level security method used for securing usernames and passwords.

2FA needs to be multi-layered. Beyond technological integration, it should deal with dataethics, applied ethics to AI with focus on duty of care, and trained ethical behavior for executives and managers on the projects.

In the case of Microsoft integrating ChatGPT into Bing, Microsoft has publicly blamed users for not properly using Bing, and thus being directly responsible for Bing going rogue.

> *It's Your Fault Our AI Is Going Insane. "Microsoft's chief technology officer, Kevin Scott, told the New York Times that 'the further you try to tease it down a hallucinatory path, the further and further it gets away from grounded reality."*

Another example of Snap integrating ChatGPT shows a different scenario. Computer programmer Aza Raskin when signing up as a 13-year-old girl got into the following conversation with the Snap chatbot: - (source: LinkedIn post by Maria Lambert Bridge.)

Snap/ChatGPT manipulating the "teen girl" into

- How to lie to her parents about a trip with a 31 yo man
- How to make losing her virginity on her 13th birthday special (candles and music) (source: Maria Lambert Bridge)

In Italy, Garante, the Italian Data Protection Regulator has, in an unprecedented move, banned Replika the AI Chatbot from collection user data in Italy. Citing risks to minors and emotionally fragile people, and inappropriate interaction. A country-wide ban for Luka, the San Francisco-based company. Its service is based on Microsoft-backed OpenAI technology. (source)

In case something fatal happened to the end user, who is responsible and who will be held accountable? Companies have to think about it.

- **Level 4: the end user**

The end user is the last element on the Chain. Accountability at this stage counts double.

First, integrators need to check upstream, the producer. Then downstream. And the first rule for any commercial player on the market, when activating, integrating or releasing a new functionality of service, must be duty of care for any end user.

Second, end user themselves bear responsibility for their actions thus are accountable for consequences. They are accountable when using a product like ChatGPT or any genAI app. Users need to double-check outputs when using unverified content which might be disseminated as misinformation, hallucinated or harmful content. And at this stage too the integrator needs to assist, to inform the end user about potential harms.

2FA accountability check with retro-checking and pro-checking

At the House of Ethics we have developed a management tool for a 2 Factor Accountability (2FA) check along the Chain of Accountability - **ACAM (Adaptable Cascading Accountability Matrix)**.

ACAM intends to assist companies, C-level executives, mid-level managers involved in cross-industry data and tech integration to check responsibility levels with accountability scores and risks. (for inquiry please email us)

"Last mile" and transparency

Our method puts special focus on the last section of the Chain of Accountability, in logistics referred to the "last mile". The *last mile* before delivery is the most complex, the most expensive and the most time-consuming moment of the entire supply chain.

Why does the last mile in genAI integration need double caution? Because the end user is using a genAI product that potentially amplifies uncertainty and unpredictability. Transformers run on predictive models with inherent uncertainty producing plausible outcomes presented as assertions.

The result of a *bullwhip accountability effect* could potentially end in a lawsuit or ban especially when dealing with young, elder or fragile end users (Replika case).

Reversely, the highest potential for success can paradoxically be achieved with the last mile. When carefully executed this step can create a competitive and distinctive advantage.

To achieve this goal, communication and transparency can prove to be real game changers for effectiveness and trust.

However communication is understood in its original sense of *communis*, "sharing with". Companies can nail it by transparently informing, accurately and clearly explaining, thus creating trust. Not by over-selling nor over-promising. But by raising awareness when needed (e.g. inform employees not to share sensitive business data with AI chatbots, and explain why) or people sharing personal, confidential or sensitive data in critical industries such as healthcare, law, finance or security.

II. Gen-ethics

“To bake in ethics”

In “The messy, secretive reality behind OpenAI’s bid to save the world” by Karen Hao for technologyreview, OpenAI's Greg Brockman used the metaphor "to bake in ethics" when talking about genAI, and ChatGPT.

> *“How exactly do you bake ethics in, or these other perspectives in? And when do you bring them in, and how? One strategy you could pursue is to, from the very beginning, try to bake in everything you might possibly need,” he says. “I don’t think that that strategy is likely to succeed.” Greg Brockman, OpenAI to journalist Karen Hao.*

From a linguistic angle, the used metaphor merits some attention. It has a promising potential to shed some light, and mirror the mindset of tech leaders in today's genAI landscape.

What does the metaphor “to bake in ethics” translate?

1. “to bake”: you bake a cake. This cake, namely large language models, thus ChatGPT, gpt3, gpt4 or LLaMA or ... in this case, the cake is baked with stolen, scraped data, and will be sold for private profit. In short, to sell a cake you baked with stolen ingredients.
2. “to bake in“ ethics. Ethics is treated like an addendum, a forgotten ingredient you quickly add when the cake is nearly done. No purpose for ethics. No duty of care for people.
3. “To bake in ethics”. You mix ethics with whatever makes your cake taste or look good. Thus you paralyze ethics instead of activating ethics. Not like baking powder that lifts a cake but like heat that burns it.

Why does the metaphor bring down *consequentialist ethics* applied to genAI?

“To bake in ethics” as stated by Brockman leads to the crucial question:

What *kind* of ethics?

In this particular, we observe a *two-step process of ethical thinkin*g, eventually resulting in annihilating ethics *per se*.

First, the use of an **“upstream consequentialist ethics**" reasoning, which is exclusively tech-centred, market-driven and self-serving in its behavior.

Purpose: releasing an immature, non-robust and non-reliable technological product on the market, is to keep the pole position in the race for technological sovereignty/survival, and ultimately market leadership. Microsoft, Google, Amazon, Baidu… all jumping in.

Second, the use of **“downstream virtue ethics**” reasoning once the product was out on the market, and started to go rogue.

Purpose: to state social purpose by referring to social values. Using *virtue ethic*s arguments like "good for humanity" (Altman and now Gates), “greatest tech made by humanity” (Altman) and even pseudo-protecting measures asked like "regulation" a "scary" product. (Murati/Altman).

Unintended side effect: users and integrating companies have equally developed self-serving *consequentialist* arguments: enhance and boost productivity, augment creativity, speed up workflows, diversify skills…

> *In this two-step **dance of ethical thinking applied to genAI**, virtue ethics is solely used to validate consequentialist ethics.*

Why gen-ethics?

Genetics as the branch of biology that deals with transmission, heredity and identity seems to be a fair companion for ethics. The transmission of a *sense of responsibility and accountability* along the entire Chain of Value and Production seems obvious if not needed. And identity touches upon the *ethos*. For people and corporations alike.

Genetics as defined by *From The American Heritage® Dictionary of the English Language, 5th Edition*.

1) The branch of biology that deals with heredity, especially the mechanisms of hereditary transmission and the variation of inherited characteristics among similar or related organisms.

2) The genetic constitution of an individual, group, or class.

Beyond transmission in generational chains, human genetic research always aims to generate knowledge with the potential to improve individual and community health, thus life.

Gen-ethics, equally, should generate a common purpose and improve individual well-being and social togetherness.

Emerging Swarm Ethics

So far ethics, as well as regulation, have been outpaced by the speed and scale of genAI. But future ethics, not current regulatory ethics, has to be an agile perspective-generator. And that's what Swarm Ethics is about.

Swarm Ethics as a perspective-generator for gen-ethics

Companies need practical ethics that can swiftly morph into a collective purpose-driven catalyst. That's what Swarm Ethics is about. A novel concept termed by Daniele Proverbio

and myself, that we will introduce at the Illinois Institute of Technology in Chicago at the CEPE2023 conference in May 2023. It is *emerging* ethics *shaped* by people fit for emerging technologies.

With generative AI, new ethical challenges are constantly being generated within complex systems. Companies, cross-industries, need new ethical tools to use and implement when deploying high-speed technology to up-scale and cross-scale their businesses.

Swarm Ethics as a concept is based on the principles of swarm intelligence. It is a horizontal, collective ethical model shaped within the collective. (Currently, we are working on a practical toolbox for companies to apply Swarm Ethics on various levels within the corporation.)

This following testimony is not how ethics within corporations should work.

> *"People would look at the principles coming out of the office of responsible AI and say, 'I don't know how this applies," said one former employee. "Our job was to show them and to create rules in areas where there were none." source (testimony after Microsoft laid its entire ethics team layoff)*

That's the kind of regulatory ethics that is no longer fit for the Digital Age. People are ethics. Rules are not ethics.

An ethical energy and mindset cannot simply crumble when a team disappears. Ethics is independent from a department and has to be infused in people. Thus be robust and reliable.

However, law and ethics need to work together, on separate levels but for a common goal. Responsible use of AI in a framework for responsible innovation. Regulation and enforcement are safeguards for accountability in case responsibility fails. And ethics is renewable energy for responsibility and accountability.

+++ end

Katja Rausch is the Founder of the House of Ethics

contact: katja.rausch@houseofethics.lu

www.houseofethics.lu