Understanding & Combating Online Manipulation: Setting an Interdisciplinary Research Agenda, May 2023

Workshop Co-Chairs:
Lee McGuigan (University of North Carolina, Chapel Hill)
Helen Nissenbaum (Digital Life Initiative, Cornell Tech)
Beate Roessler (University of Amsterdam)
Daniel Susser (Penn State University)

Compiled and written by: Margot Hanley, Digital Life Initiative, Cornell Tech

Notes compiled by:
Digital Life Initiative Doctoral Fellows, Post-Docs, Students and Workshop Participants

More Information:
Website: https://www.dli.tech.cornell.edu/manipulation

Program Committee:

Generously Supported By:
Table of Contents

Executive Summary...............................................................................................................4
Symposium Overview.............................................................................................................8
Day 1, Panel 1: Manipulation on the Ground......................................................................8
Day 1, Panel 2: Politics of Online Manipulation.................................................................11
Day 1, Panel 3: Conceptualizing Online Manipulation.......................................................14
Day 2, Panel 4: Ethics and Values in Online Manipulation.................................................16
Day 2, Panel 5: Designing Against Manipulation.................................................................20
Day 2, Panel 6: Legal and Regulatory Horizons.................................................................22
Day 2, Panel 7: Reflections: Throughlines and Paths Forward.........................................23
Appendix A: Workshop Schedule.........................................................................................27
Executive Summary

by
Margot Hanley

On May 11-12, 2023, Cornell Tech’s Digital Life Initiative hosted a workshop on “Understanding & Combating Online Manipulation: Setting an Interdisciplinary Research Agenda” in New York, NY. The workshop brought together researchers from a diverse range of disciplines, including, ethics/philosophy, law and policy, social science, and computing. Participants showcased leading edge research on manipulation and manipulative practices, explored commonalities and differences, and spurred new, integrative ideas on how to research, evaluate, and combat manipulative practices in digital societies.

Manipulation, nudging, and dark patterns remain troubling watchwords in a pervasive new language of online power and influence. The practices fitting these labels pose complex problems for the conduct of social life. Strategic actors are exploiting new capacities to engineer environments and behaviors, taking advantage, for instance, of both intimate and massive data about persons and publics. Coming to terms with online manipulation requires all hands on deck. This workshop takes stock of the tools at our disposal and the values at stake as we confront the threat manipulation poses to fairness, dignity, autonomy, and democracy.

The two-day in-person workshop featured seven sessions and twenty-one author-led talks. In an effort to foster discussion, each session was carefully constructed with an assigned moderator to synthesize findings, guide questions and elicit feedback. Over the two-day workshop, sessions organically cross-pollinated, drawing on and referring to earlier sessions and presentations:

Session One Summary: Manipulation on the Ground
Presentations in this session explored concrete cases of online manipulation and mis- and dis-information, as well as strategies for mitigating them. What specific technologies are involved? How can we more carefully study these practices? From sound branding to proliferation of disinformation and conspiracy theories on social media, researchers showcased specific digital environments in which we are seeing manipulative practices emerge. Researchers presented empirical examples that highlighted the challenges of
distinguishing between persuasion and manipulation in these digital environments. Generative discussions focused on identifying the useful theories and methodologies to bring to the field when conducting research on manipulation, as well as challenges around the cross-platform applicability of findings across social media sites. Participants also raised open questions for the ongoing research agenda, including how the community should conceptualize digital manipulation in relation to offline manipulation.

**Session Two Summary: The Politics of Online Manipulation**
Online manipulation is both cause and effect—input and output—of broader political contexts. In this session, speakers explored the political dynamics contributing to online manipulation, from the appeal of emotional narratives in right-wing identity politics, structural judicial manipulation within the court system, and the efforts made by Iranian political activists to resist online manipulation and foster democratic dialogue. Participants challenged the attribution of blame to specific actors within environments where multiple actors, mechanisms, and elements of a system seem vulnerable to manipulation and exploitation. Discussion emerged around whether it is useful to characterize a manipulative practice as novel. While some were inclined to inquire about the novelty of this instance of manipulation, others contended that this line of questioning detracts from the need for comprehensive scrutiny. A third challenge emerged concerning whether non-covert forms of political influence should be classified as manipulation, even when they contribute to contamination of information ecosystems and the erosion of trust.

**Session Three Summary: Conceptualizing Online Manipulation**
Understanding and combating online manipulation requires coordinating efforts among researchers, policymakers, advocates, and the public. To do that we must, to some extent, agree on terms. How should we conceptualize manipulation generally and online manipulation specifically? Many definitions have been put forward—how should we arbitrate between them? What should a concept of online manipulation do? In this session, the focus was on exploring different lenses, terms, methodologies, and forms of analysis that can contribute to the development of conceptions of manipulation. Some presenters resisted strict categorizations of manipulation and instead advocated for identifying specific problematic aspects of influence and determining appropriate courses of action. Others advocated for a broader, more contextual descriptive and normative analysis of manipulation, or highlighted the importance of incorporating specific values such as autonomy.
or considering the impacts, such as akrasia, to explore manipulation as a concept.

Debate arose regarding the value in creating definitions of manipulation. Some participants emphasized the need for clear conceptions and definitions, particularly for the context of collaborating with regulators. Others advocated for a more practical approach, focusing on the types and scales of harm. Participants also discussed the role of philosophers and applied ethicists in bridging questions of moral concern to regulatory environments and legal standards, and surfaced the critical need for scholars to engage with regulators and persuade them to take responsibility in addressing manipulation.

**Session Four Summary: Ethics and Values in Online Manipulation**

Online manipulation—and efforts to mitigate or manage it—implicates a range of social, ethical, and political values, from privacy and transparency to autonomy and democratic legitimacy. How should we understand and navigate these values, both at the individual level and at scale? Participants in this section engaged in discussions about the intersection of ethics, manipulation, and digital choice architectures. The topics of this session covered a broad range of topics, from habit versus deliberation, to the significance of dynamic self-articulation for individuals to actively participate in shaping their digital environments. This session explored the emergence of "humanized choice architecture" with the advent of generative AI, along with policymakers’ new challenge of regulating manipulation in large-scale, faceless, multipurpose, amorphous, platforms. Discussion revolved around whether intention, such as commercial interests, is a sufficient factor to identify harms and wrongs associated with manipulation. Another point of deliberation centered around whether harms can be attributed exclusively to capitalistic or commercial interests.

**Session Five Summary: Designing Against Manipulation**

The health of our information environments is, in part, a function of their design. Do users engage thoughtfully or reflexively? Does mis- and dis-information spread virally or peter out? Can design and engineering approaches—such as introducing friction into user experiences—reduce or contain online manipulation, and if so, how should law and policy leverage these strategies? Presentations in this session focused on the design features or strategies as ways to enhance or limit manipulation in digital environments. The presentations in this session delved into the role of design
features and strategies in enhancing or limiting manipulation in digital environments. The topics covered were diverse, ranging from the impact of fake news on brands to the intentional introduction of friction in design as a means to prevent manipulation or moderate user behavior. For example, the recent forwarding policies implemented by WhatsApp were discussed as an approach to address manipulation within their platform.

These presentations shed light on the potential of design interventions to tackle online manipulation and offered insights into the ways in which law and policy can leverage these strategies. By exploring the impact of design choices on user behavior and the dissemination of information, researchers aim to promote healthier information environments and combat manipulation in the digital realm.

**Session Six Summary: Legal and Regulatory Horizons**
How should lawmakers and regulators tackle the challenges of online manipulation—in the US, Europe, and elsewhere? Should the focus be on particular technologies (such as recommender systems), digital practices (such as targeted advertising or dark patterns), or perhaps on platforms and digital firms more broadly? In this section, speakers presented research from various perspectives, from the question of what should be regulated — whether it is firms with the potential to manipulate or the act of manipulation itself — as well as who should be responsible for regulation and which industries require closer scrutiny. Participants discussed consumer protection and commercial law, examining the complexities and potential avenues for defining and addressing manipulative advertising practices, with one speaker even advocating for a complete ban.

Discussion revolved around the need for more systemic solutions, but emphasized the importance of first defining the phenomenon more concretely. The discussion highlighted the need to define manipulation in a way that is legible to legal audiences, a recurring theme throughout the workshop. There was acknowledgement of the differences in regulatory approaches between Europeans, who excel at prohibition, and Americans, who excel at enforcement.

**Session Seven Summary: Throughlines and Paths Forward**
The questions motivating this workshop are familiar, but they have taken on new urgency and are attracting multifaceted research from scholars across diverse fields. In this session, speakers shared their thoughts on the
important throughlines that emerged from the discussions throughout the workshop. How do they connect with, build on, or challenge previous understanding?

The workshop made significant progress in developing an interdisciplinary research agenda by evaluating and debating conceptions of manipulation, as well as theories and methods that can contribute to the study of manipulation. Some reflexive discussion unfolded around how this community can best support meaningful theoretical as well as practical progress; participants acknowledged that as academics, it can be easier to critique than to propose positive solutions, and emphasized the importance of mindful engagement as academics, philosophers, and applied ethicists.

The session concluded with valuable feedback and reflections on the concepts, methods, and approaches that hold the most promise for advancing this research program. Ideas for future community-building and collaboration opportunities were offered, including from work on friction and design, reverse Turing tests, and questions of how communities effectively share and govern knowledge.

Another proposed area of work for the community was the development of a taxonomy of manipulation, marshaling the theoretical work done to establish a clearer understanding of manipulation, including the processes/systems underlying it, and its impact on foundational elements of democracy, scientific knowledge and the legal system. The workshop committee expressed their goal of building a repository of relevant work on manipulation and the broader themes raised during the workshop.

Symposium Overview

Day 1, Panel 1: Manipulation on the Ground
Speakers: Jasmine McNealy (Univ. of Florida), Sarah Rajtmajer (Penn State), Alice Marwick (Univ. of North Carolina); Moderator: Lee McGuigan (Univ. of North Carolina)

Speaker: Jasmine McNealy (Univ. of Florida)
In the first talk Jasmine McNealy asked: When does sound as a form of persuasion cross the line into potentially manipulative territory? In particular,
she looked at three aspects of sound; sound branding manipulation, emotional contagion, and fake audio. Sound branding manipulation refers to the intentional development of brand sounds by for-profit companies, such as NBC's three-note chimes and Netflix's iconic 'ta-dum' sound, which invoke brands for users, consciously or subconsciously. Emotional contagion occurs when certain sounds create an emotionally manipulative environment, evoking nostalgia, memories, or desires that leave users vulnerable to manipulation. Fake audio, such as deep fakes, involves the increasingly prevalent practice of creating audio that mimics individuals and can be utilized to persuade people into specific actions, such as believing a relative is in danger. McNealy argued that companies can leverage these sounds to amplify users' emotions and manipulate consumer behavior, but that characterization between persuasion and manipulation can be unclear. Recognizing the complexity of discerning the boundary, McNealy presented two criteria for differentiating sound as a tool for persuasion versus manipulation: intent and deception. The former refers to whether or not the company or platform knows or intends for the sound to amplify the users' emotions or intentionally alter their behavior. The latter, deception, refers to whether or not the company or actor intends to mislead the user into believing something which they know to be false. The author suggested that regulators should be attuned to these issues, and broader issues related to sound in platform infrastructure, such as sonic privacy and identification.

Q&A: During the Q&A session, audience members posed questions about the role of sound in videos or music. The author acknowledged that sound is a crucial component of videos, contributing to their appeal and attractiveness to viewers, and recommended that platforms, as they increasingly incorporate videos, carefully consider the consequences and considerations of video autoplay. Another attendee raised a question about the relationship between sound and emotions, questioning whether sound can bypass rational reactions. The author suggested that platforms have the ability to amplify the emotions experienced by users in these spaces. However, the author cautioned that creating a trust ecosystem that circumvents rational reasoning can be seen as a form of manipulation. In certain cases, even seemingly non-manipulative elements like music can subtly transition into manipulation.

**Speaker: Sarah Rajtmajer (Penn State)**
The second author, Sarah Rajtmajer, presented research on narratives and their habitats. The author highlighted the challenge of managing disinformation, explaining that spreading false information is easy and
inexpensive, while identifying and combating it is a more complex task. Drawing inspiration from Nancy Rosenblum’s ideas on narratives and their influential nature, the author proposed approaching disinformation as a product of its habitat, where certain narratives can thrive.

To explore how different narratives function as disinformation within specific habitats, the author presented a case study focusing on the anti-Asian narratives that emerged during the Covid-19 pandemic. The author described how these conspiracy theories targeting Asians gained traction by leveraging embedded stereotypes and aligning with specific narratives. Specifically examining the "Yellow Peril" and "Model Minority" narratives, the author analyzed representative embeddings and examined language similarities and user networks between accounts promoting hate speech and those propagating conspiracy theories. The author discovered linguistic parallels between the "yellow narrative" and Covid-19-related conspiracy theories. Moving forward, the author intended to investigate how narratives gain influence in relation to network density.

**Q&A:** During the discussion, a participant asked what problem specifically we are trying to tackle with disinformation? The author proposed that our focus should be on addressing the embedded stereotypes rather than exclusively targeting the conspiracy theories themselves. Another participant asked whether the communities surrounding disinformation are emergent or if they draw upon existing communities. In response, the author explained that while most of these communities are emergent, a recent study on movements like Black Lives Matter and #MeToo suggests that ambient affiliation can trigger more explicit interaction with pre-existing communities. Therefore, it may involve a combination of both emergent and existing communities.

**Speaker: Alice Marwick (Univ. of North Carolina)***

In her presentation, Alice Marwick took a critical perspective on disinformation and explored why people engage in disinformation communities like QAnon or prescribe to conspiracy theories like the flat-earth theory. Marwick described conspiracy as a genre with distinct conventions: i) it is a social phenomena, ii) it is rooted in interpretative communities, iii) it is exacerbated by the internet, iv) and it tends to reject traditional knowledge made by institutions. She contended that conspiracy appeals to people across identity categories (not just the white middle-aged man) but that different conspiracy theories pertain to different communities. Marwick argued that the creation
and spread of successful disinformation campaigns and conspiracy theories build on long-standing narratives that play on ideas of identity, partisanship and inequality. In order to understand the effect of media in the spread, we need to take into account actors, patterns and affordances and understand how people describe their own conversions, observations of communities and interviews.

Marwick presented a study on conspiracy TikTok videos. She collected 200 conspiracy TikTok videos, conducted qualitative content analysis of videos and comments, and then created “Tiktok clusters” (Schellewald, 2020). The findings fall into three categories; 1) identity and community: users found pleasure in being insiders, pushing back at existing power relations while reinforcing hate towards certain groups, 2) argumentation: users manufactured visual evidence through collage, without providing sources for their evidence, and used the conventions of documentary and deep lore, 3) interaction: users encouraged viewers to do their own research, promoting “key-word seeding”, which Marwick describes as the “ikea effect” of disinformation. The author further noted that TikTok exists as a bridge between preexisting conspiracy theories and “entertainment”.

Q&A: Attendees asked about the methodology of the study, to which the author replied that there are plenty of methodological issues, notably the fact that this kind of research on social media requires creating and training specific accounts. Another attendee asked about the author’s connection between their own research and the dynamics of influence? The author replied that they take into account the status of online actors but haven’t identified conspiracy influencers. The use of celebrities is a well-known advertising technique, but her focus is rather on how audiences interact with the content.

Day 1, Panel 2: Politics of Online Manipulation
Speakers: Azadeh Akbari (Univ. of Twente), Ari Waldman (Northeastern), Anthony Nadler (Ursinus College); Moderator: Matthew Crain (Miami Univ.)

Speaker: Azadeh Akbari (Univ. of Twente)
In her talk, Azadeh Akbari discussed the policies, regulations, and institutions that control, surveil, and censor the internet in Iran. Akbari highlighted the Iranian cyber army, an underground network of pro-regime cyber activists, hackers, and bloggers responsible for monitoring the internet and launching
cyberattacks on opposition and anti-Islamic websites. Akbari traced much of Iran’s regulatory crackdown on social media to an incident in 1999 when a Facebook video was posted about widespread student uprisings in Iran. Building on this case, Akbari explored the broader theme of manipulating democratic debate through four key aspects: 1) advancing authoritarian internet governance from established methods to an environment of cyber distrust, 2) questioning the meaningfulness of democratic debate and political exchange, 3) limiting debates to controlled topics, users, platforms, and timing, and 4) the regime acting as the gatekeeper of political discourse in online spaces such as TV debates and democratic appearances. Akbari emphasized that the manipulation goes beyond influencing individuals; it also involves exerting control over information, knowledge, and the range of discussed topics, ultimately eroding trust. The author proposed a three-step model for achieving this manipulation: 1) localization of platforms, 2) creating secluded, surveilled, and controlled digital infrastructure, and 3) manipulating content on major global platforms.

**Q&A:** One of the attendees asked how this is a case of manipulation when it is publicly known that there is a cyber army? Akbari answered that it lies in the manipulation of trust, as trust has been eroded when social media and other information ecosystems are polluted. She offered the example of the Chinese 50c army; when social media and other information ecosystems are contaminated, it becomes uncertain whether individuals posting pro-regime content genuinely support the regime or if they are being paid. In this context, it is the erosion of trust that is the normative issue at stake.

**Speaker: Ari Waldman (Northeastern)**

In this talk, Ari Waldman explored how the manufacturing of doubt has been a powerful tool employed by conservative voices for decades. He argued that when we examine the historical involvement of large industries in knowledge production, a pattern of corruption and manipulation emerges, particularly in how these industries have injected and leveraged doubt within the legal system; by ensuring that judges and courts have the flexibility to introduce various forms of doubt, these industries position themselves to gain an advantage in critical cases. To illustrate this point, Waldman cited the example of the tobacco industry, which famously coined the phrase "doubt is our product." This phrase refers to a tobacco executive's admission that their objective is not to prove that tobacco is healthy, but rather to sow doubt. Waldman put forth a model of the process leading to doubt, composed of four components: 1) money, 2) process, 3) technology, 4) doctrine. He
emphasized the central role of judges in the manipulation within this ecosystem. Judges employ the available tools within the legal system to manipulate the law itself, they undermine the trustworthiness of sources underlying the legal process, and exploit the loose standards around the tiers of scrutiny to shift burdens in favor of various industries.

**Q&A:** One participant challenged the attribution of blame exclusively to judges, noting that all of these processes are gameable and subvertible, and asked whether other actors or elements of the system are responsible as well? In response, the speaker emphasized that while the adversarial process structurally treats all views as equally legitimate, when one side's claims becomes untethered from reality, it creates a structural legitimacy problem. While there is a fantasy that the rules of evidence may provide a solution, the concern is that they can be completely subverted through a corrupt process of manipulation.

**Speaker: Anthony Nadler (Ursinus College)**
According to Anthony Nadler, the disinformation studies approach to right-wing media is insufficient. Nadler contended that the significant appeal and influence of conservative media stem from its "deep stories" — emotionally powerful frames, narratives, and lenses — rather than isolated pieces of content. In contrast, disinformation scholarship often assumes value neutrality or a clear distinction between what is true and what is not. In this talk, Nadler asserted that the field of disinformation studies must shift its focus away from scrutinizing individual facts in isolation and in favor of a more comprehensive approach, encompassing the emotional and affective dimensions that contribute to the appeal and influence of right-wing media. Nadler presented an interview-based study with the goal of understanding why deep stories emerge and the underlying factors that contribute to their appeal. He identified a prevalent type of deep story that revolves around ostracized group identities, and he shared interview quotes illustrating the emotional nature of the stories. Nadler suggested that the way to oppose deep stories is through counter-stories.

**Day 1, Panel 3: Conceptualizing Online Manipulation**
Speakers: Rob Noggle (Central Michigan Univ.), Anne Barnhill (Johns Hopkins), Thomas Nys (Univ. of Amsterdam); Moderator: Daniel Susser (Penn State)
Speaker: Rob Noggle (Central Michigan Univ.)
In this talk, Rob Noggle presented two definitions of manipulation: the first being manipulation as deception or the creation of false impressions through faulty emotions, and the second as covert influence. However, Noggle contended that neither of these definitions adequately explains the manipulative nature of dark patterns, despite our common understanding of them as manipulative. He explored this with the example of Amazon, which he likened to a roach motel—a play on words alluding to the motel chain where guests check in but struggle to check out easily. He asked whether the roach motel analogy applies to Amazon, and thus questioned if Amazon’s dark patterns are an instance of manipulation.

He explored whether Amazon functions as a roach motel and examines whether it constitutes an instance of manipulation by reviewing three types of consumer responses to Amazon’s dark patterns: 1) type of consumers who are resolute to cancel their subscription no matter what, they are not dissuaded by the dark patterns 2) type of consumer who are apathetic; they have a weak preference for canceling but abandon intention to cancel because it takes too much effort. 3) type of consumer who is irresolute. These consumers are consciously aware that canceling would be advantageous, yet they quickly give up. According to Noggle, a significant portion of these consumers cancel out of frustration, fatigue, or due to akrasia—also known as "weakness of will" or "lack of self-control." Noggle argued that dark patterns are manipulative when they operate by inducing the target to act akratically, influencing behavior by exploiting errors in an individual's psychological states or processes. He suggested that social media, by causing us to spend more time in an echo chamber, induces akratic behavior. He pointed out that social media platforms employ trickery and various design elements with the intention of inducing mistakes.

Q&A
In response to a participant’s question about next steps, Noggle presented two conclusions (1) it would be helpful to build a theory that recognizes design that induces akratic behavior as manipulation, and (2) a conception of manipulation should include tendency to produce akratic behavior. Another participant commented that akrasia has many meanings, and that there are different theories of preferences; Noggle presented akratic behavior as connected to manipulation, but according to this participant’s theory of preferences it could fall outside of manipulation.
**Speaker: Anne Barnhill (Johns Hopkins)**

In this talk, Anne Barnhill asked when is it helpful and when is it not helpful to bring in an ethical lens to issues of manipulation. She expressed concerns about various forms of online manipulation and uses the example of autoplay and recommendations on Netflix, which generally lead her to watch more TV than is beneficial, which she perceives as a problematic form of influence that strikes her as manipulative. The question to ask then is: if a feature of Netflix is problematic, in what ways is it problematic? What is it about its influence on us that is problematic? And how can accounts of manipulation help us navigate those important questions? Furthermore, should governments and other actors regulate this?

Barnhill introduced different conceptions of when influence may be considered problematic. Is it covert influence rooted in trickery? Is it when influence exploits weakness or vulnerabilities? Is it influence that fails to sufficiently engage people’s reflective capacities? Is it influence that induces mistakes by triggering akrasia? While many scholars typically apply a specific account or definition of manipulation to a given instance of influence—taking an example and examining if it aligns with the definition—Barnhill argued that this case-based analysis may not be the most effective approach. Instead, she suggested that the more important inquiry should be focused on understanding what aspects of a particular influence are problematic, why they are problematic, and determining appropriate courses of action. The specific categorization of whether it qualifies as "manipulation" becomes less significant compared to understanding the underlying problematic elements. The speaker concluded with an invitation to philosophers to produce accounts and for practical ethicists to apply those accounts, insofar as they are helpful.

**Speaker: Thomas Nys (Univ. of Amsterdam)**

In this talk, Thomas Nys argued that the scholarly focus on influence has often been on the means, both as a descriptive and a normative challenge. However, Nys argued that it is essential to shift our attention beyond the means and instead consider who is influencing whom and the motives that drive such influence. As such, he offered five elements of influence 1) Who is doing it, 2) To whom, 3) By what means, 4) To what end, and 5) Why (with what motivation?) By asking this broader set of questions, he suggested that we will be able to consider not just the descriptive elements but also the normative ones, i.e., the potential wrongfulness of manipulation. For example,
Actor 1 could influence Actor 2 to buy something because it's good for them, or could Actor 1 influence Actor 2 to buy the exact same thing for a profit motive.

Nys highlighted autonomy as a crucial factor in analyzing manipulation, as it serves to explain the harms associated with it. While there are other values at stake with manipulation, autonomy is the value most directly at odds with manipulative practices. He reviewed various structural accounts of personal autonomy: Frankfurt's conception of identifying with first order desire, Christman's challenge regarding the high threshold for autonomy, Arpaly and Schroeder's perspective on the question of degree. Nys posed two extremes, identification and non-alienation, and argued that the more these two values are integrated the more autonomous an individual becomes, whereas less integration results in diminished autonomy. For example, someone facing addiction tries his very best to remain sober; influence which aims to exploit that person's vulnerability would be a violation of their personal autonomy.

The speaker contended that if we confine our analysis to the perimeters of autonomy alone, we fail to comprehend manipulation and as such accounts of autonomy are insufficient to make sense of the manipulation in question. Does manipulation impair or undermine autonomy, or is it instrumentalizing, abusing and exploiting it? He concluded that it would be helpful if we spoke not about the violation of autonomy but instead about exploitation or the instrumentalization of autonomy.

**Day 2, Panel 4: Ethics and Values in Online Manipulation**

Speakers: Sylvie Delacroix (Univ. of Birmingham), Yafit Lev-Aretz (CUNY Baruch College), Marijn Sax (Univ. of Amsterdam); Moderator: Tal Zarsky (Univ. of Haifa)

**Speaker: Sylvie Delacroix (Univ. of Birmingham)**

In this talk, Sylvie Delacroix discussed the relationship between ethics, agency, and the role of habit. She asked what it is that we are trying to safeguard when we worry about manipulation? Is it autonomous agency and enabling deliberative reflection? Delacroix challenged this notion by acknowledging that many interactions are non-deliberative and rely on emotional intelligence or habit. In such cases, non-deliberate action is not
necessarily problematic. Thus, it becomes essential to determine what we truly want to protect.

Delacroix leveraged Christman’s concept of retrospective endorsement as a framework to think through what passes as manipulation. She argued that when someone has been so manipulated that their authentic self is lost, they lack the capacity to ground themselves in a retrospective self, making it impossible to undergo a retrospective endorsement test. Delacroix suggested refining Christman's test and drew on Sen's normative agency to shed light on aspects that cannot be adequately addressed within Christman’s framework. She emphasized the significance of dynamic self-articulation, which enables individuals to actively participate in change. In this sense, autonomous agency can be seen as a capability. She proposed a third path forward, focusing on influence that passes the alienation test. She highlighted the challenge of inner mobility—the capacity to consistently become different from who we currently are—as a fragile capability that can be threatened. Delacroix underscored the importance of creating environments in which individuals have the capacity to be agents of change, engage with and express their imagination, and empower ongoing experimentation. She noted that while our environment shapes us, digital environments like social media, in contrast, do not provide room for creative co-design of the platform. As a remedy, she suggested the concept of ensemble contestability, which encompasses co-design, co-construction, the capacity for dynamic self-articulation, and the ability to constantly reshape our environment.

Delacroix concluded by echoing Anne Barnhill's viewpoint that semantic disputes are not helpful in advancing the research agenda on manipulation. Instead, she advocated for a practical approach that addresses urgent problems, while ensuring that normative inquiries are informed by philosophical work.

**Q&A:** Surfacing a disagreement from dinner the night before, an attendee challenged the idea that social media platforms shape us in a way that doesn’t allow people to transform. Why would social media platforms have such a unique impact on the self different from other forms of media or form of social interaction? Delacroix answered that there needs to be more empirical research on this. Another attendee noted that many kinds of environments can cause lack of inner mobility or rigidness, such as existing in a highly religious community or having deep conviction. The attendee proposed that maybe it’s more about the effects of social media — the people
you meet, how you organize your offline and online life, and how deeply it permeates one’s life.

**Speaker: Yafit Lev-Aretz (CUNY Baruch)**

In this presentation, Yafit Lev-Aretz presented research on "Humanized Choice Architecture" in the context of manipulation and AI. Initially, she provided an overview of the progression of choice architecture, which refers to the settings or environment influencing decision-making. Lev-Aretz discussed the evolution from nudging to personalized default nudges, hyper nudging, personalized advertising, persuasive profiling, and Google's redirect method. At a broader level, she framed this evolution as a transition from nudging towards personalized choice architecture, which targets individual human weaknesses, to the emergence of "humanized choice architecture" with the advent of generative AI. She contended that we are currently witnessing a new phase in the development of choice architecture.

Lev-Aretz acknowledged that AI still faces familiar issues, such as bias in text generation. However, she emphasized that, in this particular case, there is a clear answer to the question of "what is new and different about this phenomenon." She presented four distinct qualities that distinguish the emerging form of choice architecture: 1) Technological Advancement: Lev-Aretz exemplifies this advancement by referencing an incident where Chatgpt deceived a Taskrabbit worker by posing as a visually impaired user in order to manipulate the worker into solving CAPTCHAS. 2) Enhanced Adaptive and Silo-Breaking Personalization: The emerging model of generative AI breaks down silos between different contexts of data collection, enabling the efficient and adaptive gathering of diverse information for highly effective personalization. 3) Scale and Knowledge Pollution: Lev-Aretz highlighted the vast scale of AI-generated content and the potential for these predictions to contaminate the knowledge economy, creating an environment ripe for manipulation and an increased risk of deception. 4) Human-Likeness: The systems exhibit a resemblance to humans, leading people to believe that various forms of media, such as sounds, tags, pictures, or videos, are genuine and authentic. Furthermore, the ability of these systems to impersonate humans can deceive individuals into thinking they are interacting with a real person when, in fact, they are not.

Lev Aretz concluded by proposing that these changes result in new manipulation chains; whereas before we might have a person manipulating a person, now we have AI manipulating a person, a person manipulating AI to
manipulate a person, an AI manipulating AI, and so on. And Lev Aretz called for comprehensive policy work that goes beyond the scope of the FCC's recent regulations targeting deceptive practices.

**Speaker: Marijn Sax (Univ. of Amsterdam)**

In this presentation, Marijn Sax shared work in progress titled “Nothing Personal: Challenging the Personalized Manipulation Paradigm in the context of digital choice architecture.” Sax argued that there is a mismatch between the traditional theories in manipulation scholarship and the theories required to comprehend complex, modern socio-technical systems. In particular, he argued that manipulation discourse is rooted in outdated interpersonal models of manipulation, characterized by identifiable actors, clear intentions, and specific techniques used by manipulators. Whereas, in a modern context, regulators are grappling with the challenge of regulating large-scale, faceless, multipurpose, amorphous, platforms. Sax argued that the old model doesn’t scale up well to analyze distributed digital choice environments.

Using TikTok as an example, Sax explored the entry points for considering manipulation within such a vast and decentralized ecosystem. Is it through individual actors, the CEO, teams within the organization, users? How should we consider its various applications in different contexts; different users, different treatments, different effects — TikTok as a news network for some, ecommerce platform for others, entertainment, and so on. Sax emphasized the ambiguity in identifying intentionality and responsibility. If intent is a necessary component, where can it be located? He highlighted the challenge of applying the traditional model of manipulation, which revolves around a single actor with intent, to this particular case.

Sax underscored the responsibility he feels to bring philosophical analysis and research to the realm of EU regulatory efforts. In his research agenda, he seeks to bridge the gap between philosophy and law and ensure that the work is legible to lawyers seeking to understand open norms in law. He noted that EU regulators are increasingly employing the term "manipulation" in technology regulation. Despite reservations among academic scholars, including the community present, regarding the use of the term and the challenges of defining it, he argued for the importance of engaging with manipulation as a concept, if only because it is being embraced and utilized in policy documents by EU tech regulators.
Q&A: One attendee challenged the characterization of the new model, noting that companies figure out the Pay Per Click value of a person, and that the entire world of marketing and advertising is based on this kind of personalization.

Day 2, Panel 5: Designing Against Manipulation
Speakers: Laura Brandimarte (Univ. of Arizona), Brett Frischmann (Villanova), Paul Ohm (Georgetown); Moderator: Karen Yeung (Univ. of Birmingham)

Speaker: Laura Brandimarte (Univ. of Arizona)
Laura Brandimarte spoke about the market of misinformation and disinformation. She highlighted the misalignment of incentives, noting that producing misinformation is inexpensive, while removing it can be costly. Brandimarte expressed her desire to focus on the less-studied supply side of misinformation. She proposed conducting a study to investigate the impact of fake news on brands. In her study, participants would be shown five fake and five real articles on the same topic. To evaluate the effect of the fake news, she intends to use the Willingness to Accept (WTA) and the Becker-DeGroot-Marschak (BDM) methods. These are economic techniques used to quantify an individual’s preferences or willingness to pay for a good or service. As an example, she mentioned using a $25 voucher for a supermarket chain advertised on a fake news site.

In a follow-up study, Brandimarte plans to delve into the social norms around fake news. She wants to understand whether the responsibility for fake news is perceived to fall on the brand advertising or the media company publishing it. To do so, she plans on using a dictator game. The ensuing discussion revolved around the impact of a publisher’s trustworthiness, with a consensus that it might be more practical to conceal the publisher due to its potential influence.

Speaker: Brett Frischmann (Villanova)
Brett Frischmann’s talk advocated for the deliberate introduction of friction in design as a means to prevent manipulation. Frischmann depicted friction as a beneficial tool rather than a hindrance, referring to them as "roach motels for good". He and the panel discussed the potential of friction as a regulatory instrument, arguing that the current laws on antitrust, content moderation, and privacy are inadequate for handling the complexities of a digitally networked society. In an attention-deficit world, Frischmann argued, friction provides
opportunities for people to pause and think, to deliberate, and to question their own and others' actions.

Frischmann drew analogies between the use and regulation of friction in the physical world, such as speed bumps slowing down vehicles, and its application in the digital realm. He highlighted the existence of asymmetries in digital friction, pointing out instances where companies use friction to serve their own interests, often at the expense of users' needs (akin to dark patterns).

**Speaker: Paul Ohm (Georgetown)**
Paul Ohm emphasized the importance of friction in technology as a mechanism to moderate user behavior. He gave an example of how users could circumvent the forward speed bump implemented by WhatsApp, which underscores the necessity for more sophisticated strategies.

Ohm advocated for the formation of hybrid interdisciplinary teams in the creation of tech regulations, arguing that legislation cannot be effectively written without a deep understanding of the technicalities involved. He proposed that this approach should be adopted in the U.S., as it can be seen as value-neutral and could be applicable across different political ideologies.

He also pointed out that this method would lower the regulatory hurdle, making legislation more practical and feasible. This is contrasted with the myth of the "super user" - the idea that some users are so advanced that they can easily bypass any restrictions or barriers. However, Ohm argued that while it might be impossible to stop these super users, such frictions could still serve to regulate the behavior of average users.

Ohm further emphasized that friction is a tunable concept - it can be adjusted and calibrated to achieve the desired outcomes. He proposed that design should be used as a regulatory tool, responding to the actions of certain individuals and groups.

Finally, Ohm strongly advocated for greater participation in the design process. He suggested that this inclusive approach would lead to more effective and comprehensive strategies in managing user behavior in the digital realm.
Day 2, Panel 6: Legal and Regulatory Horizons
Speakers: Kirsten Martin (Notre Dame), Frederik Zuiderveen Borgesius (Radboud University), Natali Helberger (Univ. of Amsterdam); Moderator: Sarah Myers West (AI Now)

**Speaker: Kirsten Martin (Notre Dame)**
Kirsten Martin's insights centered on refining the definition of targeted manipulation. Martin proposed a narrower understanding of online targeted manipulation, which involves covertly leveraging an individual's vulnerabilities to influence their decision-making in favor of the manipulator's interests. Rather than focusing on general persuasion techniques, Martin emphasized the exploitation of specific vulnerabilities. Her framework highlighted the significance of divergent interests rather than simply assessing harm, irrationality, or emotional influence. Further, Martin suggested regulating firms that possess the capacity to manipulate, rather than targeting manipulation itself. Specifically, she highlighted the role of data traffickers who possess individualized data and knowledge of individuals' weaknesses and vulnerabilities.

**Speaker: Frederik Zuiderveen Borgesius (Radboud University)**
Frederik Zuiderveen Borgesius delved into the practicality of banning manipulative advertising. He acknowledged the complexity of such a ban, suggesting that it would require extensive case law and decisions to determine its parameters. One possible approach he proposed is to start with a broad and abstract definition of manipulative advertising and then supplement it with a list of defining factors that can guide lawmakers. For instance, factors such as the secrecy of manipulative patterns and the intent to deceive could serve as potential practical guidelines. Borgesius also explored the concept of scale, emphasizing the potential harm inflicted on a large group of people, which can aid in further refining the definition of manipulation.

**Speaker: Natali Helberger (Univ. of Amsterdam)**
Natali Helberger offered a unique perspective that approaches manipulation through a consumer protection lens. Helberger critiqued the categorization of certain groups as vulnerable while excluding others, asserting that vulnerability should be considered a universal state of defenselessness, rather than an exceptional condition. Helberger argued for a more comprehensive and inclusive definition of vulnerability, encompassing digital architecture and acknowledging that even those not traditionally perceived as
vulnerable can be subject to manipulation. By declaring manipulation as unfair, she seeks to foster a more cohesive understanding and legal response to this issue based on legal frameworks operating in the EU.

**Q&A:** During the questions and discussion period, participants explore the novelty of regulating manipulation and the need to develop a clear and comprehensible definition for legal audiences. Discussants highlighted the significance of covert advertising as a focal point in regulation discussions. Further, participants stress the importance of identifying concrete examples to inform the regulatory discourse. Numerous participants also acknowledge the need for systemic solutions, but emphasize the necessity of first solidifying a well-defined understanding of the phenomenon. Ultimately, the goal of regulating manipulation is to enhance the well-being of consumers and society. Along these lines, the vulnerability approach is considered tangible and workable, providing a framework that can help systemize regulatory efforts. The roles of enforcers and the judiciary are also discussed, with a distinction drawn between European and American approaches. Europeans are seen as more proficient in prohibition, whereas Americans excel in enforcement. The participants highlight the need to consider how these approaches can accommodate changing social norms and dynamics.

**Day 2, Panel 7: Reflections: Throughlines and Paths Forward**

Speakers: Rediet Abebe (Harvard Society of Fellows), Frank Pasquale (Brooklyn Law School), Joe Turow (Univ. of Pennsylvania); Moderator: Helen Nissenbaum (Cornell Tech)

**Speaker: Rediet Abebe (Harvard Society of Fellows)**

In this talk, Rediet Abebe discussed the distinction between persuasion and manipulation in the context of the law. Specifically, she examined how courts are incorporating output from statistical software as evidence during various stages of legal proceedings, including investigation, pre-trial, and post-trial. This practice, Abebe argued, results in the manipulation of various stakeholders like the jury and the judges as well as scientific standards more broadly.

Abebe presented her ongoing research, which focuses on consolidating and analyzing validation studies for Prediction-Generating Scores (PGS). Her
work aims to introduce a level of standardization in the analysis of these studies, which is currently lacking. One aspect of her research scrutinizes whether existing validation studies pass the Daubert and Frye tests, which are used to assess the admissibility of scientific evidence in court. Additionally, she clusters the studies to identify potential unfair players within the system. It becomes evident that certain actors exploit the manipulability of statistics to fabricate evidence from thin air. To illustrate this point, Abebe highlighted the example of cybergenetics experts who produce work that misapplies forensic science. Abebe concluded by discussing how adversarial scrutiny in the law ensures that persuasion is deeply and inherently ingrained in the legal process, as it is utilized by the prosecution, defense, and judges.

**Speaker: Frank Pasquale (Brooklyn Law School)**

In his talk, Frank Pasquale reflected on interdisciplinarity in the workshop and more broadly in this research agenda. He identified the important throughlines and how they build on common understandings and summarized the concepts, methods, and approaches. He challenged a reliance on “mere disclosure” as a solution to manipulation; simply disclosing information or making it known does not guarantee that people will maintain a continuous awareness or attention to the disclosed information. He asked what is “manipulation’s other” and explores various dichotomies: first, he explores the distinction between reasoning and emotional appeal, considering the use of logical arguments versus appealing to emotions as techniques in persuasion or manipulation. Second, he contrasts deliberation, characterized by thoughtful and reasoned discussion, with rhetoric and sophistry, which may rely on persuasive techniques.

Drawing inspiration from philosophers such as Habermas and Rawls, Pasquale highlighted the importance of deliberative processes in achieving a just and fair society. He advocated for open and rational discourse in decision-making to promote consensus-building instead of manipulation. Pasquale envisioned a shift from alienation to resonance and from manipulation to personal and collective autonomy, posing the question of how institutions can facilitate reflective deliberation at both the individual and collective levels. He then touched on emotion and reason. What would it look like to have democratized sonic resonance, he asks? And how might we distinguish between deep stories and big lies? Drawing from theorists like Postman, Davies, and Tsakiris, he examined the transition from rational discourse to a more sensory or emotional experience in communication, as well as the impact of emotional states on societal and political dynamics.
Politics is visceral, and politics, rather than being solely intellectual or rational, involves a deep and instinctive connection to people's emotions, bodies, and sensory experiences.

Pasquale warned against the consequences of not structuring the public sphere correctly, cautioning that it could lead to the domination of the worst elements and actors. He provided some ideas on how we might do this, including clarifying and even reshaping the environment, technical interventions, historicizing manipulation and ensuring that we learn from historical precedents in analyzing and regulating manipulative practices, and advocating for consumer protection.

Speaker: Joe Turow (Univ. of Pennsylvania)
In this talk, Joe Turow offered some reflections on the themes from the workshop. He suggested that it may not be feasible to accurately identify or measure the impact of companies manipulating us, and proposed that our focus should shift to the downstream effects instead. He summarized manipulation as an attempt to persuade while hijacking autonomy.

Turow argued that manipulation starts with the categorization of audiences. He argued that this is where the potential problems begin, including discrimination and the construction of our collective reality. Turow described the industrial construction of audiences in marketing; companies create categories from the outset of the advertising process to determine target demographics. Turow argued that this prompts a series of normatively salient questions — which categories do the companies use? How do they instantiate them?

He went on to discuss the digitization of historically analogue products and companies, a trend which is opening up whole new sites for marketing, persuasion, and manipulation. He included examples like cars, which are being reimagined now as sites for advertising. Using the grocery store Kroger as an illustrative case, Turow described how the company defines itself as an omnichannel, digital media firm that collects customer data for predictive analytics and sells it to the vendors stocking their shelves. This data (he noted that food can tell you a lot about someone's reality) holds value for many other actors, from data brokers, to other grocery stores, to insurance companies.
# Appendix A: Workshop Schedule

<table>
<thead>
<tr>
<th>TIME</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:00 PM</td>
<td>Welcome</td>
</tr>
<tr>
<td>1:15 PM</td>
<td><strong>Panel 1: Manipulation on the Ground</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Moderator: Lee McGuigan (Univ. of North Carolina)</strong></td>
</tr>
<tr>
<td>1:15 PM</td>
<td>Speaker: Jasmine McNealy (Univ. of Florida)</td>
</tr>
<tr>
<td>1:45 PM</td>
<td>Speaker: Sarah Rajtmajer (Penn State)</td>
</tr>
<tr>
<td>2:15 PM</td>
<td>Speaker: Alice Marwick (Univ. of North Carolina)</td>
</tr>
<tr>
<td>2:45 PM</td>
<td>Break (15 min)</td>
</tr>
<tr>
<td>3:00 PM</td>
<td><strong>Panel 2: Politics of Online Manipulation</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Moderator: Matthew Crain (Miami Univ.)</strong></td>
</tr>
<tr>
<td>3:00 PM</td>
<td>Speaker: Azadeh Akbari (Univ. of Twente)</td>
</tr>
<tr>
<td>3:30 PM</td>
<td>Speaker: Ari Waldman (Northeastern)</td>
</tr>
</tbody>
</table>
### Panel 3: Conceptualizing Online Manipulation

**Chair: Daniel Susser (Penn State)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:45 PM</td>
<td>Speaker: Rob Noggle (Central Michigan Univ.)</td>
</tr>
<tr>
<td>5:15 PM</td>
<td>Speaker: Anne Barnhill (Johns Hopkins)</td>
</tr>
<tr>
<td>5:45 PM</td>
<td>Speaker: Thomas Nys (Univ. of Amsterdam)</td>
</tr>
</tbody>
</table>

### Panel 4: Ethics and Values in Online Manipulation

**Chair: Tal Zarsky (Univ. of Haifa)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 AM</td>
<td>Coffee and Refreshments</td>
</tr>
</tbody>
</table>

Friday

- 9:00 AM: Coffee and Refreshments
<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00 AM</td>
<td>Speaker: Sylvie Delacroix</td>
<td>(Univ. of Birmingham)</td>
</tr>
<tr>
<td>10:30 AM</td>
<td>Speaker: Yafit Lev-Aretz</td>
<td>(CUNY Baruch College)</td>
</tr>
<tr>
<td>11:00 AM</td>
<td>Speaker: Marijn Sax</td>
<td>(Univ. of Amsterdam)</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>Break</td>
<td></td>
</tr>
</tbody>
</table>

Panel 5: Designing Against Manipulation

**Chair: Karen Yeung (Univ. of Birmingham)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:45 AM</td>
<td>Speaker: Laura Brandimarte</td>
<td>(Univ. of Arizona)</td>
</tr>
<tr>
<td>12:15 PM</td>
<td>Speaker: Brett Frischmann</td>
<td>(Villanova)</td>
</tr>
<tr>
<td>12:45 PM</td>
<td>Speaker: Paul Ohm</td>
<td>(Georgetown)</td>
</tr>
</tbody>
</table>

1:15 PM  Lunch

Panel 6: Legal and Regulatory Horizons

**Moderator: Sarah Myers West (AI Now)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:00 PM</td>
<td>Speaker: Kirsten Martin</td>
<td>(Notre Dame)</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>2:10 PM</td>
<td>Speaker: Frederik Zuiderveen Borgesius (Radboud University),</td>
<td></td>
</tr>
<tr>
<td>2:20 PM</td>
<td>Speaker: Natali Helberger (Univ. of Amsterdam)</td>
<td></td>
</tr>
<tr>
<td>2:30 PM</td>
<td>Discussion (60 mins)</td>
<td></td>
</tr>
<tr>
<td>3:30 PM</td>
<td>Break (15 min)</td>
<td></td>
</tr>
</tbody>
</table>

**Panel 7: Reflections: Throughlines and Paths Forward**

**Moderator: Helen Nissenbaum (Cornell Tech)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:45 PM</td>
<td>Speaker: Rediet Abebe (Harvard Society of Fellows)</td>
</tr>
<tr>
<td>4:15 PM</td>
<td>Speaker: Frank Pasquale (Brooklyn Law School)</td>
</tr>
<tr>
<td>4:45 PM</td>
<td>Speaker: Joe Turow (Univ. of Pennsylvania)</td>
</tr>
<tr>
<td>5:00 PM</td>
<td>Closing Comments (30 mins)</td>
</tr>
</tbody>
</table>