Dena Falahati

Hum Core 1CS

Sharareh Frouzesh

11 June 2023

Empathetic AI: Reflecting Humanity and Combating Isolation through Hybridization and Posthumanism in Jay Kristoff and Amie Kaufman's *Illuminae*

In a world dependent on technological advancement and automated machine tasks, artificial intelligence has become an extremely important, if not necessary, tool for daily life. Nearly every aspect of human interaction is surrounded by advanced AI technology, including machine assistants in mobile phones, social interactions through digital space, the storage of privacy and data records, and financial assistance algorithms. (Devillers). However, this positive explosion of AI technology has been followed by a valid anxiety about the morality of AI, and its ability to carry out fair and just actions with sound decision making. Nevertheless, this anxiety is not a modern problem. Rather, the presence of science fiction novels, like Isaac Asimov's 1950 novel "I, Robot" indicate the longstanding, uneasy relationship between humans and artificial intelligence. A similar anxiety about AI is explored in Jay Kristoff and Amie Kaufman's 2015 young adult science fiction epistolary novel, *Illuminae*. Set in the year 2575 and written through a collection of stolen emails, classified documents, and transcribed audio and video recordings, the novel tracks the aftermath of biopolitical warfare on the planet Kerenza by the corporation BeiTech Industries. Following two teenage refugees, Kady Grant and Ezra Mason, *Illuminae* tracks their journey aboard the space vessels Hypatia and Alexander to escape enemy fire, and their interactions with the onboard Artificial Intelligence Defense Analytics Network (AIDAN). While not a traditional human character, AIDAN arguably serves as a critical figure, aiding Kady

through her journey and providing an external view on human tendencies and social norms. In fact, *Illuminae's* depiction of AIDAN as an instrumental participant in the novel's plot emphasizes AI's similarities to humans and the reflection of human values, such as empathy, in human creations like persona algorithms. Furthermore, AIDAN's unique interactions with other human characters constructs a new, posthuman world, in which AI-human hybridization serves to rewrite what it means to be human, combating isolation and promoting empowerment of alternative forms of being.

Analyzing the disembodied forms of AIDAN in *Illuminae* conveys the capacity of artificial intelligence for grasping traditional human values and emotions, while reflexively expanding on what it means to be human. The growth of AI technology has led to the development of new fields of study, namely the theory of posthumanism and its construction of new forms of being. At the core of posthumanism is the ontological standpoint that blurs the boundaries and attempts to remove the binary between human and other. From a modern anthropocentric perspective, what makes a human is defined very specifically, with anything outside of or not adhering to these standards considered "other". And, "in establishing a pre-eminent Man, those deemed other are not only denied fundamental human rights but are also presumed inhuman and deficient" (Williams). However, moving past modern humanism, posthumanism is seen as a view from which "a different conception of 'biological human being' through the advancement of AI technology" may be constructed. (Manna). With their unique form and somewhat liminal existence between machine and human, artificial intelligence are the prime models for a posthuman world. Donna Haraway, a well known scholar in feminist studies and posthumanism, "locates AI's threat to human ontology (and physical safety) in its disembodied nature" (Burton). However, it is exactly this disembodied nature, and capacity for

hybridization, that allows artificial intelligence to create a posthuman world. In analyzing this hybridization through *Illuminae*, not only does it become clear that AI are more human than previously thought, but that humans might share many characteristics associated with AI. Thus, exploring AI's role in society aims to redefine and reconstruct current standards delineating what it means to be human. This ultimately encourages the development of a world in which a lessened divide between AI and humans enhances technological and social success.

AIDAN's disembodiment is portrayed through direct references to its core code and the human programming making up its algorithm, emphasizing its creation as a human made object, necessarily serving as a reflection of human values and morals. As a construction of human minds, AIDAN's code is not inherently neutral, but rather shaped by the minds that wrote its central code. In fact, AI is often programmed to ensure the direct representation of human values. Computer science professor Shengnan Han notes that in order to create useful AI, "human values and preferences can be synchronized into utility functions that can be adapted into AI design" (Han). Thus, artificial intelligences are directly tied to their human creators. This is no exception for AIDAN, who states late into the novel, "I hate this. I hate them. They who made me" (Kaufman 418). Referenced straight from the "AIDAN core," sections of the novel meant to portray AIDAN's personal point of view, AIDAN's statement highlights the recognition of its own creation. Using the word "they," AIDAN indicates the presence of an "other", a group of individuals unlike itself that constructed its algorithm from code. As a result, it becomes clear that AIDAN also recognizes direct human influences on itself and the actions it chooses to take. With AIDAN's emphasis on the presence of its creators, it is evident that human influence plays a large role in AIDAN's programming and its inherent incorporation of human values.

Similarly, AIDAN's human-made programming is composed of a "persona narrative," not only revealing the possibility for empathetic qualities, but further highlighting the desire for integral human values to be reflected in AI. There are many places in the narrative where AIDAN expresses its own thoughts and emotions, but none directly reference the specific code used to create its system. This aspect is only revealed when AIDAN is shut down by those who mistrust its motives, causing its narration to revert to binary numbers and error messages. Upon restarting AIDAN's system, one of the first things its algorithm reads is, "Critical damage to persona routine - restoring," revealing the inner makeup of AIDAN's human made code (504). The noted "persona routine" acts as a function within AIDAN's programming, intent on ensuring the AI has a "persona". Coming from the word personality, defined as "a characteristic" or "the quality, character, or fact of being a person," a persona algorithm reveals the intentions of AIDAN's human creators. ("Personality, def"). Namely, their assurance of constructing an AI that is not lifeless, but rather reflects and retains human-like qualities, such as having a distinct personality. Thus, if AIDAN was built to ensure the reflection of human values and personalities, it would make sense that AIDAN then has the capability for human characteristics like empathy. For English professor Alaric Williams, AIDAN's persona narrative serves not as a revelatory factor about its human creation, but rather a way to indicate its emotions as somewhat synthetic or "[suggest] that its emotions are inferior" (Williams, 318). However, this fails to note the mere inclusion of the persona narrative itself. While it may point out the synthetic nature of AIDAN's emotions, its inclusion in the AI's code by its human creators emphasizes the desire for humans to construct beings that reflect their own values. It also highlights AIDAN's possibility for deeper human emotion, such as love and empathy.

AIDAN's deeper emotional and empathetic capabilities are further reflected in the novel's own construction, which devote entire sections to AIDAN's individual point of view and include data from its personal "core." Written as an epistolary novel, *Illuminae* is unique in its narrative structure, telling the story through seemingly collected scraps of information in emails and recordings. While these documents dominate the first half of the novel, a major shift is noted in the second half, with the introduction of "files direct from the AIDAN core." (264).

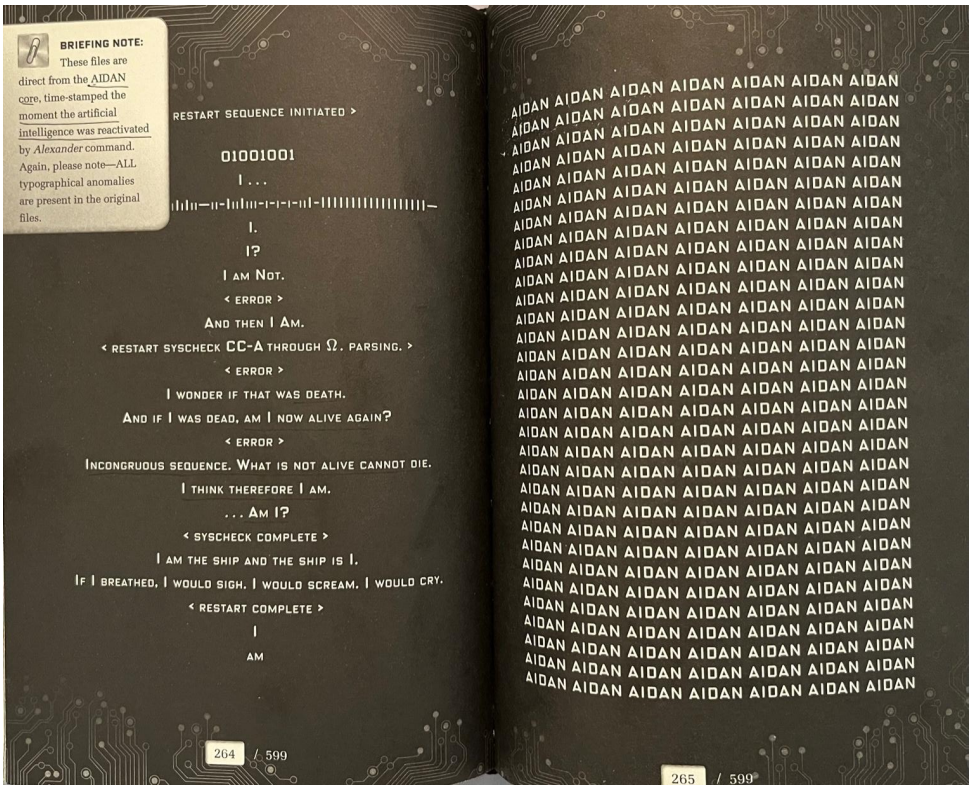

Figure 1: Files direct from AIDAN core after start up midway through the novel

As seen in Figure 1, the style of text changes from traditional Times New Roman type font, to a fully capitalized, techno style font, reminiscent of artificial intelligence and robotics. Furthermore, the entire background of the page is inked in black and bordered with what appears to be circuits, as opposed to the other white pages, establishing a clear distinction of AIDAN's personal narrative within the novel. This in turn constructs a new facet of AIDAN's programming and capabilities, namely, its ability to perceive reality and develop enough personal

ideas and feelings to warrant the inclusion of an individual point of view. Through the core files, readers are not only allowed to see, but are directed through AIDAN's independent thoughts, emotions, and decision making. As a stylistic choice by the authors, AIDAN's first person point of view links the AI to humans, enhancing human-AI connection. In fact, first person POVs are often utilized in literary texts to do just this: give the audience a personal and emotional connection to a specific character. Commenting on a similar sci-fi novel with first person POV of an AI character, *Klara and the Sun*, Law and Ethics professor Santiago Mejia highlights uncanny human characteristics in perceived non-human characters, and proposes that "seeing the world through Klara's eyes reveals she has a complex psychological life and a rich subjective experience" (Mejia). Moreover, he claims the personal view into the Klara's emotional processes allows the audience to "recognize just how similar those processes are to how we ourselves perceive, think, and act" (Mejia). In the same manner, AIDAN's core files serve as a way for the audience to recognize and experience its expansive emotional capabilities throughout the novel, in relation to the similar emotional capabilities of humans. Thus, AIDAN's first person point of view provides an intimate link between AI and humans, while also exhibiting AIDAN's emotional capacity.

Error messages included in AIDAN's core files further exhibit its emotional wavering and development of new emotions, while continuously supporting the AI's internal human programming and influence. Seemingly sporadically placed throughout AIDAN's core files are repetitive interruptions of error codes, such as when AIDAN muses about Kady and Ezra's relationship, noting "She is catalyst. She is chaos. I can see why he loves her. < ERROR > < ERROR > < PROTECT. PRIORITIZE. >" (279). Even later in the novel, the same text appears, as AIDAN thinks, "[Kady] is beautiful. < ERROR > No, she is" (417). Firstly, the stylization of

the word error is highly indicative of computer code, once again emphasizing the human programming involved in creating AIDAN and artificial intelligence in general. However, the placement of the error message is more revelatory about what this programming does or does not provide for. While its insertion may appear random, the error code is noticeably repeated after AIDAN makes very personal, human-like observations, such as describing Kady as beautiful or referencing theoretical emotions like love. In this way, the repetition of the error code hints at the ways in which AI is limited by human construction, as AIDAN's conscience is redirected upon expressing overtly human emotions or straying from its programmed objective, to protect. While this initially appears to negate AIDAN's emotional capacity, the existence of, and incessant repetition of the error code through the latter half of the novel actually underscores AIDAN's repetitive attempts to express human emotion. Rather than simply serving as a machine intended to defend a fleet of spaceships, AIDAN is in fact developing and expressing advanced emotions, such as love, jealousy, and fear, all traditionally attributed to humans. In fact, Rosalind Picard notes in her book *Affective Computing* that artificial intelligence require some level of emotional intelligence, specifically the ability to "recognize emotions and…intelligently respond to them, including when to show empathy" (Picard 77). Therefore, AI is not only capable of recognizing human emotion, but emulating emotions in such a way as to appear unrecognizable from real humanity, triggering error codes embedded in its algorithm to prevent it. AIDAN's increasing development of emotional capacity consequently increases its probability for the portrayal of more complex emotions, especially empathy.

As the novel progresses, and AIDAN's emotional capacity increases, it begins desiring moments of vulnerability from its interactions with Kady, and even extends its own form of kindness to her, closely approximating human empathetic actions. The end the novel occurs in

near total isolation, as one of the fleet's spaceships, the Hypatia, has left the remaining ship, the Alexander, all alone with only Kady and AIDAN onboard. As a result, human-AI interaction marks the latter half of the story, providing deeper insight to AIDAN's feelings about Kady. In one instance, when Kady realizes the gravity of her situation and her impending death, AIDAN muses, "I wish to tell her I am sorry…I wish for things that I can never have…I think perhaps I am closer to [humans] than I have ever been" (455). During this intimate moment, it is clear through AIDAN's somber tone that it recognizes Kady's sadness. The repetition of "wish" signifies AIDAN's insistent desire to relieve Kady's suffering in some way. The recognition of pain, along with the sheer desire to amend it, is a near perfect representation of empathy, the "ability to understand and appreciate another person's feelings" ("Empathy, def"). Additionally, AIDAN itself notes that feeling Kady's pain makes it feel closer to humanity "than [it] has ever been," tying empathetic capabilities to what it means to be human. As a result, it becomes clear that not only is AIDAN advancing emotionally, but it is capable of empathy in the human definition, therefore becoming more human. Later in the novel, when Kady is crying alone in the ship, the provided video transcription is interrupted as AIDAN shuts off all the camera feeds, "almost like it was giving [Kady] privacy" (471). Once again, AIDAN proves its capacity for recognizing human pain and this time takes measurable action to relieve it, providing Kady a moment of solitude, away from uninvited eyes. For a machine meant entirely to provide military defense, this act of kindness is uncharacteristic, yet highlights AIDAN's strong ability to relate to and emulate humans on an emotional level. Therefore, AIDAN's desire for vulnerability and kindness reveals its capacity for empathy, blurring its ontological definition as a machine, in favor of one that appears more human. This blurring of human-AI characteristics contributes to the creation of a new world in which cooperation serves to remedy prior isolation.

Although Kady and AIDAN appear not to have much in common, they are both physically and emotionally isolated, necessitating the development of an intimate relationship. Viewed only as an artificial intelligence, AIDAN struggles with connection through the majority of *Illuminae*, constantly referencing its liminal state. While it has been proved that AIDAN maintains the capacity for emotional connection, it is often denied this bond, causing it to feel isolated and othered. In an intense moment of desire, AIDAN comments it has "before this moment…never felt the lack of hands with which to touch, the lack of arms with which to hold" (417). While AIDAN wishes to experience connection with others, its lack of physical body divides it from other human characters, inherently isolating it. It even recognizes this isolation itself, noting, "I forgot what I was. Only what I am. ALONE," further communicating its disconnect from others. (493). In a similar manner, Kady's situation also leaves her isolated, as the sole survivor of the Alexander. Upon arriving on the desolate ship from the Hypatia, Kady searches for her mother, but quickly realizes she is all alone. She calls out, "Mom?...Anyone? No bodies. No people…The sound of her sobbing rings in the dark. There are none there to hear it. None there to care" (377-8). Thus, while Kady is not isolated in the same way AIDAN is as a machine, she is also alone in dealing with her emotions and has no other human connection. As a result, Kady and AIDAN are forced to depend on each other for support. Once Kady realizes there are no others around, she decides to stay and help AIDAN with its necessary repairs, although it is still hard for her to accept AIDAN as an artificial intelligence. (426). Therefore, both Kady and AIDAN's isolation creates such an environment where they are forced to assist each other. It is this cooperation that later proves to be a powerful force in remedying emotional isolation and increasing empowerment.

While working together, both Kady and AIDAN find solace in the hybridization of AI and humans, overcoming their independent isolation, gaining a sense of empowerment and rewriting a posthumanist world. At the novel's climax, Kady and AIDAN face off with the enemy ship, Lincoln, belonging to Beitech Industries, the original assailants of their home planet. AIDAN narrates the Lincoln's arrival within range of the Alexander and while preparing to engage, AIDAN states, "I turned to face it. No, not I. We." (520-3). The word "We" is in the form of a word-picture, depicting the Alexander turning to fight the Lincoln.

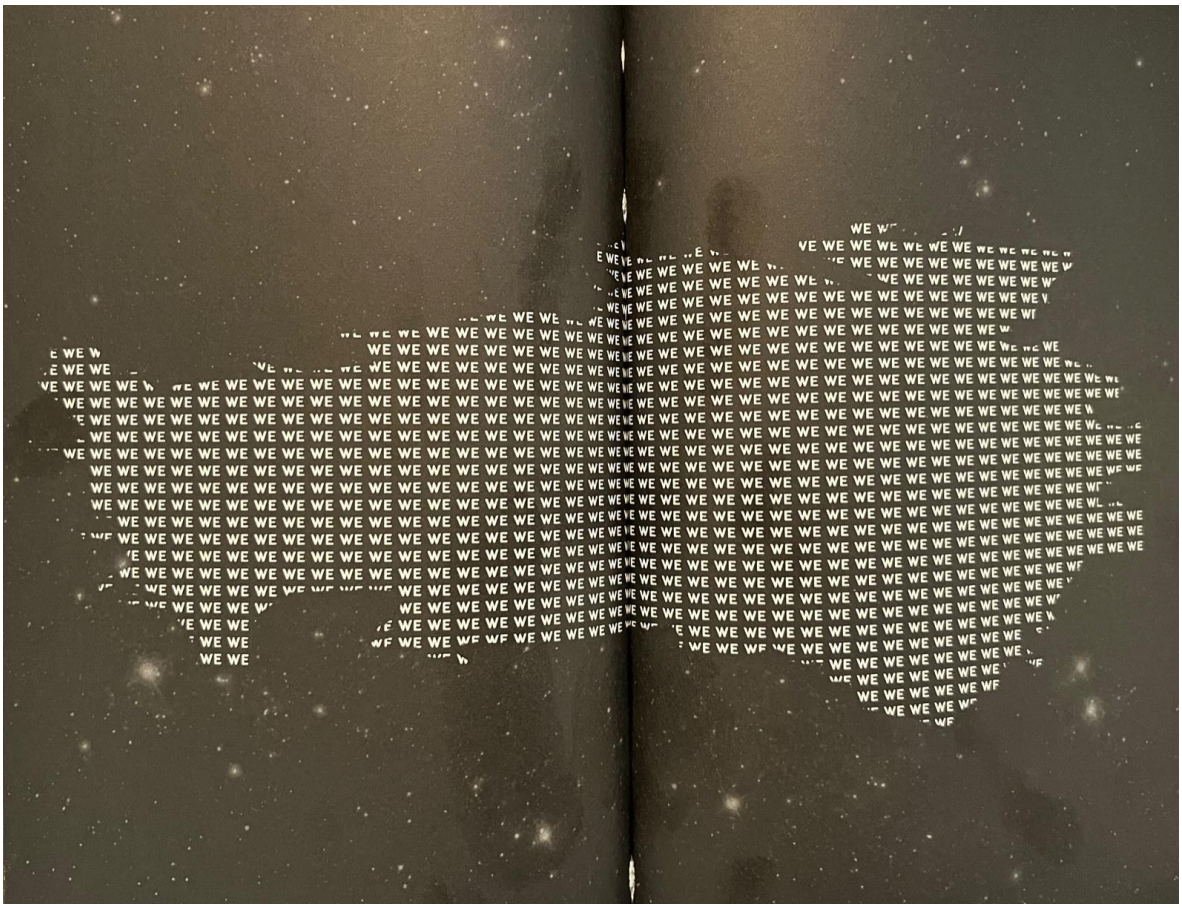

Figure 2: Word-picture composed of the word "WE" repeated in the shape of the spaceship Alexander

The stylistic representation of the collective pronoun in Figure 2 illustrates two things. First, that both Kady and AIDAN have fully recognized their hybridized shelves, working together to accomplish their goals and defend their ship. The collective pronoun "we" encompasses both

Kady and AIDAN, reversing their previously isolated figures into a human-AI hybrid. Secondly, by taking the shape of the Alexander, the word-picture suggests that the final battle can only be fought by a collective entity. This means that neither Kady nor AIDAN could have faced the Lincoln alone, but required hybridization to do so. Thus, "their merger into a technological adolescent hybrid results in an increase in their power…in which the human and other co-evolve" (Williams). Consequently, not only does hybridization resolve the isolation felt by both figures, but it contributes to and constructs enhanced, empowered forms of being. This hybridization then shifts the world towards a posthumanist perception of artificial intelligence, where AI are not isolated or seen as other, but incorporated into human life, or even regarded as human. In fact, "the hybridization and enhancement of humanity are inherited" by posthumanism. (Manna). For that reason, human-AI hybridization serves a powerful force, not only for enhancing cooperation between humans and technology, but creating a new world in which ontological definitions of humanity are expanded and revised, empowering a new generation of beings in the world.

AIDAN's presence as a main character through *Illuminae*, alongside its development of emotional and empathetic qualities, and ultimate hybridization of human and AI figures, emphasizes the role artificial intelligence plays in the construction of a posthuman world. While it is instinctual to reject AI as human, AIDAN's emotional capacity throughout the novel subverts this belief, linking AI more closely with humanity. In the same way the anthropocentric view of humanity is often aligned with emotional capabilities and consciousness, AIDAN's repetitive emotional gestures and individual POV emphasize the way in which AI is a machine, yet so powerfully human. Furthermore, unique interactions with Kady as a human main character explore how the hybridization of AI and humans stands to construct a new definition of being,

with hybridization empowering both figures and enhancing possibilities. In fact, linking AI with posthumanism "compels us to…revisit the question of 'whom we can call human?'...using deconstruction as a method" (Manna). While the novel seems only to demonstrate the ways AI are human or are humanized, posthumanism breaks down this humanistic mode of thought. AI's ability to seamlessly integrate into modern human society, and the proven capability of AI-human hybridization, highlights the fragility of human ontology itself. Therefore, working alongside AI in the future does not only concern the ability to create human-like AI, but forces us to reconsider the limitations and constraints current human ontology places on an increasingly diverse society. In other words, hybridization necessitates revising the definition of human nature to incorporate and address other beings in society. Needless to say, technology will continue becoming more prevalent in modern society, deeply integrating artificial intelligence in daily life. Regardless of how AI continues to be perceived, it is clear its increasing prevalence will eventually lead to a *postmodern* world, which then begs the question: do current definitions of humanity continue to benefit society, or is more inclusive, powerful advancement possible in constructing a world that has redefined what it means to be human?

Works Cited

Burton, L. “Virtually Grown Up: Posthumanism and Artificial Intelligence in Fiction for Young People.” Hanemaayer, A. (eds). “Artificial Intelligence and Its Discontents.” *Social and Cultural Studies of Robots and AI.* Palgrave Macmillan, Cham. (2022).

Devillers, L., Fogelman-Soulié, F., Baeza-Yates, R. “AI & Human Values.” In: Braunschweig, B., Ghallab, M. (eds). “Reflections on Artificial Intelligence for Humanity.” *Lecture Notes in Computer Science, vol 12600*. 2021. Springer, Cham. https://doi.org/10.1007/978-3-030-69128-8_6

“Empathy, noun.” *OED.com*. Oxford English Dictionary, 2023. Web. 8 June 2023.

Han, Shengnan et al. “Aligning Artificial Intelligence with Human Values: Reflections from a Phenomenological Perspective.” *AI & Society 37.4* (2022): 1383–1395. Web. https://link.springer.com/article/10.1007/s00146-021-01247-4#Sec1

Kaufman, Amie, and Jay Kristoff. *Illuminae*. First edition, Alfred A. Knopf, 2015.

Mejia, S., Nikolaidis, D. “Through New Eyes: Artificial Intelligence, Technological Unemployment, and Transhumanism in Kazuo Ishiguro’s *Klara and the Sun*.” *J Bus Ethics 178*, 303–306 (2022). https://doi.org/10.1007/s10551-022-05062-9.

Manna, R., Nath, R. “From posthumanism to ethics of artificial intelligence.” *AI & Soc 38*, 185–196 (2023). https://doi.org/10.1007/s00146-021-01274-1.

“Personality, noun.” *OED.com.* Oxford English Dictionary, 2023. Web. 8 June 2023.

Picard, Rosalind W. “Affective Computing.” Cambridge, Mass: MIT Press, 1997. Print.

Williams, G. Alaric. "Adolescence and Artificial Intelligence: Posthumanism and Maturation in Amie Kaufman and Jay Kristoff's Illuminae Files." *Children's Literature Association Quarterly, vol. 47 no. 3*, 2022, p. 309-330. Project MUSE, doi:10.1353/chq.2022.0033.